

Article

Patent Keyword Analysis Using Bayesian Factor Analysis and Social Network Visualization in Digital Therapy Technology

Sunghae Jun 

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

Abstract: Patent keyword analysis involves extracting and examining keywords from patent documents. Since patents contain detailed information about technologies, analyzing them can provide valuable insights for efficient technology management. This paper proposes a novel method for patent keyword analysis that integrates Bayesian factor analysis and social network visualization. Using Bayesian factor analysis, we extract factors representing key technologies within a domain. These factors are used as nodes in a social network analysis to explore their interrelationships. The resulting visualization aids in constructing a technology diagram, enhancing understanding of the technological structure. To evaluate the method, we analyze patents related to digital therapeutic technologies. Experimental results demonstrate the approach's effectiveness and its applicability to practical technology analysis tasks. Our results indicate that data analysis serves as a core technology in the field of digital therapy, while technologies such as remote patient monitoring, device systems, and signal processing function as supporting technologies for data analysis. The findings contribute to technology management practices, including strategy development, by analyzing target technologies and deriving actionable insights across various domains, including digital therapeutics.

Keywords: Bayesian factor analysis; keyword analysis; patent data; social network visualization; technology analysis



Academic Editor: Hovhannes
Harutyunyan

Received: 18 January 2025

Revised: 16 February 2025

Accepted: 19 February 2025

Published: 20 February 2025

Citation: Jun, S. Patent Keyword Analysis Using Bayesian Factor Analysis and Social Network Visualization in Digital Therapy Technology. *Computers* **2025**, *14*, 78. <https://doi.org/10.3390/computers14030078>

Copyright: © 2025 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

To understand technology, patent data analysis is usually performed. Patent documents provide various and detailed information about developed technologies, including the title, abstract, citations, claims, technology codes, application and registration dates, inventors, and drawings [1,2]. Among them, we focus on keywords extracted from patent documents and analyzing them. To facilitate data analysis using statistical and machine learning methods, it is necessary to transform the patent keyword data into a structured format [3–5]. Through preprocessing with text mining techniques, we construct a patent-keyword matrix as structured data. In this study, patent keyword analysis is conducted to analyze the technology keywords extracted from patent documents related to the target technology. This approach enables a comprehensive understanding of the target technology through the analysis results. Previous research on patent keyword analysis has primarily modeled correlations or associations between the patent keywords [6–9]. Most of these studies have relied on statistical methods such as regression and classification or visualization based on social and Bayesian networks. However, for a more precise understanding of technology, it is essential to uncover not only the relationships between keywords but also the latent structures among them.

Therefore, we propose a patent keyword analysis method utilizing Bayesian factor analysis (BFA) to identify the latent associative structure of patent keywords. BFA is a method that applies Bayesian inference to traditional factor analysis (FA) [10]. In order to identify the latent structure between patent keywords, we derive analysis results based on the posterior distribution that cannot be confirmed through existing FA by considering uncertainty and prior information. In our study, we use the covariance structure of patent keywords to identify latent variables, called factors, that explain the unobserved but hidden technological association structure of the entire patent data. Traditional FA extracts factors based on a linear combination of observed variables (keywords), but BFA defines factors based on a prior distribution and a likelihood function based on observed data [10,11]. That is, we model the technological uncertainty between patent keywords that describe the target technology using BFA to understand the nonexposed association structure between detailed technologies. In BFA, the prior distribution of factor loadings is reflected in the observed data and is ultimately estimated as the posterior distribution for factor loadings. In this process, the Markov Chain Monte Carlo (MCMC) sampling method is used. The BFA results allow us to extract factors that represent detailed technological components, which are subsequently visualized through social network analysis. Finally, our technology diagram for target technology is constructed by visualization of the results, closeness, and betweenness measures of social network analysis (SNA) [12,13]. To demonstrate how our proposed approach can be modeled and implemented in practice, we collect and analyze patent documents related to digital therapy technology. Using our method, we estimate the posterior distributions of factor loadings and factors representing various sub-technologies for digital therapy, and carry out SNA visualization. Consequently, the technology diagram of digital therapy is built to understand this technology.

This paper is structured in the following sections. In Section 2, we deal with our research background, such as FA, social network visualization, literature review, and patent text mining. We show the proposed method in Section 3. In Section 4, we carry out experiments using practical patent documents related to digital therapy to show how this paper can be applied to practical areas. Then, in Section 5, we discuss the results and several implications of our study. Lastly, we illustrate the conclusions, contributions, limitations, and future works in the Section 6.

2. Related Works

2.1. Factor Analysis

FA is conducted to represent the correlated variables as the latent variables called factors [14–16]. The idea of FA is to describe the p variables $(X_1, X_2, \dots, X_p)$ by m factors $(f_1, f_2, \dots, f_m)$, where m is much smaller than p . Using the results of FA, we can achieve a better understanding of the related p variables. The general FA model is defined as follows [11]:

$$X_i = \mu_i + a_{i1}f_1 + a_{i2}f_2 + \dots + a_{im}f_m + e_i, \quad i = 1, 2, \dots, p \quad (1)$$

In Equation (1), a_i is the factor loading of the i th variable on the j th factor, and e_i is the error representing the variation of X_i not explained by the factors. The observable random vector X has the mean μ . In this paper, X is the frequency value of the patent keywords and f is a technology defined by the patent keywords. We also use Bayesian inference to improve the performance of FA.

2.2. Social Network Analysis

SNA (social network analysis) is a process to find the social structures of nodes (variables) using graph theory [12,13]. A graph data structure represents the data nodes by the directed and undirected connections between the nodes [17,18]. The graph is defined

as $G(V, E)$, where V is the node (vertex), and E is the edge. That is, V and E are the data objects and the connections between the data objects in the graph data structure, respectively [12,17]. Therefore, the SNA is conducted to analyze the relational data and find the social connections included in the observed data using the graph structure. To date, SNA has been studied and utilized in a wide variety of fields, such as biology, economics, geography, social psychology, technology management, etc. [12]. In the field of technology management, SNA has been used as an efficient analysis method to identify the relationship structure between technologies [12]. In this paper, we apply FA based on Bayesian inference to SNA for visualizing the patent keywords. From the social network visualization, we build a technology diagram to understand the target technology.

2.3. Research Background and Literature Review

Patent analysis has been performed in various ways for technology management purposes. Feng et al. (2020) studied patent analysis using morphology analysis and unified structured inventive thinking to discover technology opportunities [19]. They used text-mining techniques and Word2Vec clustering to find the intrinsic connections between innovation elements in the management of technology. In addition, Hu et al. (2025) carried out a patent analysis related to the technology of CO₂ reduction [20]. The authors applied bibliometric analysis to their technology analysis. Using the results of their research, they found that the advanced manufacturing, research, and technology of catalytic materials are continuously needed for CO₂ reduction. Yang et al. (2025) used social network analysis to construct a patent network for studying technology communities and dominant technology lock-in in the Internet of Things field [21]. This research constructed a directed citation network using 9464 IoT patent families as nodes and 23,604 inter-patent citation relationships as directed links.

2.4. Patent Text Mining

The goal of text mining is to transform unstructured document data into a structured document-term matrix. Figure 1 shows the collected documents and the document-term matrix constructed by text mining.

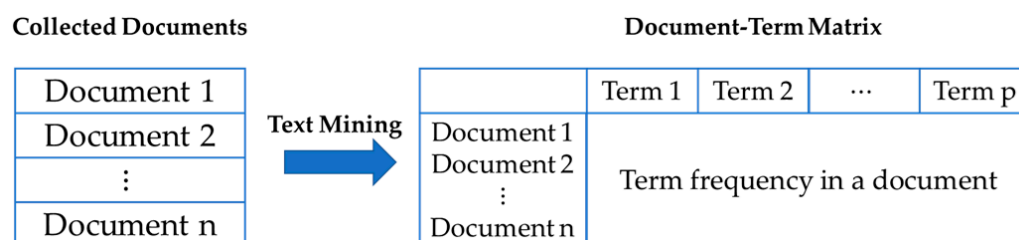


Figure 1. From collected documents to document-term matrix.

A document-term matrix is a structured representation of text data where rows represent documents and columns represent unique terms across all the documents. The cells of the matrix contain the frequency of each term in each document. In this paper, the patent keywords are selected from the terms. Thus, the keyword text mining procedure is performed as follows [22–26]:

Step 1: Document collection

- (1-1) Searching documents related to the target domain;
- (1-1) Filtering valid documents.

Step 2: Text preprocessing

- (2-1) Converting texts to lowercase;

- (2-2) Deleting punctuation and stop words;
- (2-3) Stemming;
- (2-4) Removing special characters.

Step 3: Tokenization

- (3-1) Selecting the type of tokenization as a word;
- (3-2) Splitting texts into tokens (terms).

Step 4: Definition of text databases

- (4-1) Creating a list of all terms;
- (4-2) Counting the frequency values of each term occurring in each document.

Step 5: Construction of document-term matrix

- (5-1) Initializing the matrix with (number of documents $\times$ number of terms);
- (5-2) Building the matrix using text databases;
- (5-3) Extracting keywords from the terms.

In general, the collected documents are transformed into the document-term matrix through the procedures given in Step 1 to Step 5.

3. Proposed Method

Patents contain detailed results of the developed technology and the scope of rights to the technology. Thus, we can understand the technology of the target domain by analyzing the patents related to the target technology. In particular, the analysis of technology keywords extracted from patent documents is a good way for us to effectively examine the target technology. In this paper, we propose a method of patent keyword analysis for technology management. Our method consists of two steps. The first step is a preprocessing process to build a structured patent keyword database, and the second step is an analysis process to analyze the structured database using our proposed method.

3.1. Preprocessing for Patent Keyword Data

The patents include not only the title and abstract, but also comprehensive details about the developed technology, such as the inventors, registration date, claims, drawings, citation information, and technology classification codes [1,2]. Among them, we use the title and abstract for patent keyword analysis. We search the patent documents related to the target field from patent databases across the world [1]. The collected patent documents have to be transformed into a structured database for patent analysis [3,5,7]. In this step, we build a structured keyword database using text-mining techniques. Figure 2 shows the preprocessing of patent documents using text mining.

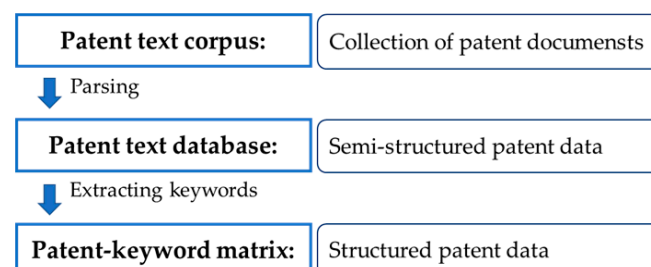


Figure 2. Preprocessing of patent document data.

We construct a patent text corpus using the searched patent documents. This is a collection of patent documents. Using the preprocessing of grammatical parsing, the text corpus is transformed into a semi-structured patent database called a text database. Lastly, we construct a patent-keyword matrix as a structured database by extracting technology

keywords from the patent text database. This matrix consists of patent documents and technology keywords for its rows and columns respectively. Each element of the matrix represents the frequency of a keyword contained in a patent document. We analyze this matrix data using the proposed method explained in the following sections.

3.2. Patent Keyword Data Analysis Method Combining Bayesian Factor Analysis and Social Network Visualization

BFA is a multivariate statistical analysis method that extends traditional FA to the Bayesian statistical approach [11,27]. BFA uses prior distribution, likelihood function, and posterior distribution in the process of estimating latent variables to explain the correlation structure between the observed variables representing the keywords. Through this, we can estimate factors hidden in related data better than general FA models and deal with uncertainty better. That is, conventional FA, which estimates parameters using the principal component and maximum likelihood methods, has difficulty clearly explaining the probabilistic uncertainty hidden in the data [11]. The uncertainty about latent factors is updated each time through the prior distribution and the likelihood function for the observed data. In the BFA, the roles of the prior, likelihood, and posterior distributions are shown in Table 1.

Table 1. Role descriptions of prior, likelihood, and posterior distributions on BFA.

Distribution	Role Description
Prior	Setting initial beliefs about parameters by reflecting prior knowledge about the distribution of variables and factors
Likelihood	Model explanation using observed data
Posterior	Update parameter values by combining prior distribution and likelihood function

By Bayes' rule, the posterior probability $P(\theta|X)$ is expressed as the following Equation (2) [10,28,29]:

$$P(\theta|X) = \frac{P(\theta)P(X|\theta)}{P(X)} \quad (2)$$

where θ is the parameter and $P(X)$ is calculated as following Equation (3).

$$P(X) = \int P(\theta, X)d\theta = \int P(\theta)P(X|\theta)d\theta \quad (3)$$

Since $P(X)$ is independent of the model parameter θ and is generally difficult to calculate, the posterior distribution is obtained as follows:

$$P(\theta|X) \propto P(\theta)P(X|\theta) \quad (4)$$

In Equation (4), we obtain the posterior distribution by multiplying the prior distribution $P(\theta)$ and the likelihood function $P(X|\theta)$. To analyze the patent keyword data, we use the BFA, which applies Bayesian inference to FA, in this paper. Based on the existing FA model described in Equation (1), we represent the BFA model as follows [11,27]:

$$X_{(p \times 1)} = \mu_{(p \times 1)} + A_{(p \times m)}F_{(m \times 1)} + \varepsilon_{(p \times 1)} \quad (5)$$

In Equation (5), $A \sim N(0, \Sigma_A)$, $F \sim N(0, I_m)$, and $\varepsilon \sim N(0, \sigma_2 I_p)$. The observed variable X consists of mean vector μ , factor loading A , factor F , and error vector ε . The factor loading matrix A follows a normal distribution with mean 0 and variance–covariance matrix Σ_A . The latent vector F follows a normal distribution, with mean 0 and a unit vector

variance of size m . Finally, the mean and variance of the error vector ε , which follows a normal distribution, are 0 and σ_2 , respectively. Also, in this paper, we choose the inverse gamma distribution with shape parameter α and scale parameter β as the distribution of σ_2 . To estimate the parameters, the samples are drawn from the posterior probability distribution. Since accurately obtaining the posterior distribution is generally challenging, the MCMC method is employed. In this study, the Metropolis–Hastings algorithm is adopted as the MCMC technique. This algorithm constructs a Markov chain that explores the search space to generate samples from the target posterior distribution. It achieves this by repeatedly proposing new candidate samples and determining their acceptance based on the acceptance probability [10,11,27]. From the results of the BFA, we find the factors and their keyword lists with factor loading values.

We use factor loadings to select keywords corresponding to each factor. In Table 2, m and p_i represent the number of factors and the number of keywords included in factor i . The keywords are used to define the technology that each factor represents. For example, the technology represented by Factor 1 is defined by p_1 keywords (Keyword 1, Keyword 2, ..., Keyword p_1). As with general FA, BFA requires determining the optimal number of factors before starting the regular analysis. In general, there are several methods to determine the optimal number of factors in FA [14,15]. Among them, we use the criterion of eigenvalues to determine the optimal number of factors. The number is determined by the number of factors with eigenvalues greater than 1. Therefore, in our study, we perform social network visualization using factors greater than 1.

Table 2. Keyword list with factor loadings.

Factor	Keyword List
Factor 1	Keyword 1, Keyword 2, ..., Keyword p_1
Factor 2	Keyword 1, Keyword 2, ..., Keyword p_2
$\vdots$	$\vdots$
Factor m	Keyword 1, Keyword 2, ..., Keyword p_m

Next, we carry out social network visualization using the factors representing technologies. SNA is an analysis method that models the social relationships and interactions between nodes in a network. In this paper, the nodes correspond to factors representing technologies. SNA is based on graph theory, which consists of nodes and edges. Using the visualization results of the SNA graph, we construct a technology diagram that describes the relationship structure between technologies. In this paper, we use the correlation coefficients between nodes (keywords) to visualize the social networks [12,13]. First, a correlation coefficient matrix between each keyword is constructed from the patent-keyword matrix. Next, a new matrix is created in which each correlation coefficient value of this matrix becomes 1 if it is greater than or equal to a given threshold value, and 0 if it is less than the threshold value. The results of social network visualization vary depending on how this value is determined. As this value increases, the social network structure becomes more complex. Conversely, beyond a certain point, a higher value simplifies the network structure, making it more challenging to interpret the associations between keywords. In this study, we analyzed the relative network structure between keywords by adjusting the threshold value and ultimately selected an appropriate threshold. Finally, this matrix consisting of 0 or 1 values is used as an adjacency matrix for social network visualization. Figure 3 shows the process of social network visualization.

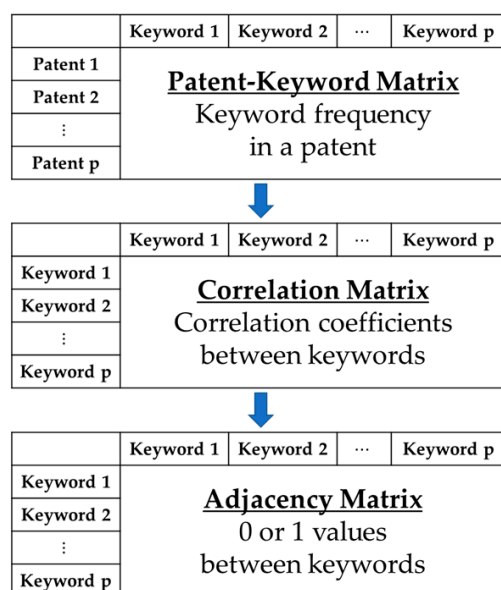


Figure 3. Process of social network visualization.

We build the correlation matrix from the patent-keyword matrix and make the adjacency matrix from the correlation matrix. In social network visualization, if the value of the adjacency matrix is 1, the two corresponding keywords are connected, and if it is 0, they are not connected. In this process, we use three measures of SNA: degree, closeness, and betweenness. In SNA, degree represents the number of edges a node has. This metric measures the connectivity of nodes within a network. The degree of node i (d_i) is defined as in Equation (6) [13]:

$$d_i = \sum_j A_{ij} \quad (6)$$

where A_{ij} has the value 1 if nodes i and j are connected, and 0 otherwise. The larger the degree value of a node, the greater its influence in the network. The closeness in SNA is a measure of how close a node is to all the other nodes in the network. The closeness of a node i ($C(i)$) is defined as follows [13]:

$$C(i) = \frac{1}{\sum_{i \neq j} s(i, j)} \quad (7)$$

In Equation (7), $s(i, j)$ represents the shortest path between nodes i and j . The higher the closeness value of a node, the faster the node can transmit information to other nodes in the network. Finally, the betweenness in SNA is a metric that measures how many shortest paths a node has between pairs of nodes in the network. The betweenness of node i is defined as the sum of the number of shortest paths passing through i divided by the number of all shortest paths in the network, as shown in the following Equation (8) [13]:

$$B(i) = \sum_{h \neq i \neq j} \frac{N_{hj}(i)}{N_{hj}} \quad (8)$$

where, N_{hj} is the total number of shortest paths between nodes h and j , and $N_{hj}(i)$ is the number of shortest paths between nodes h and j that pass through node i . We use degree, closeness, and betweenness to find the technology structure for the target domain based on the social relationships between each factor from the SNA visualization results. We call this a technology diagram. Figure 4 shows the entire process of our proposed method.

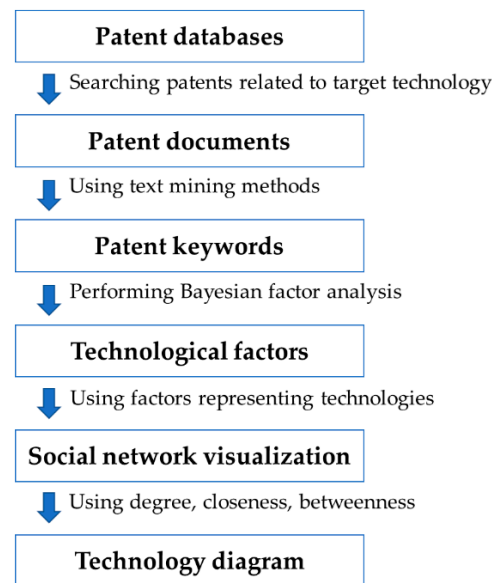


Figure 4. Process of the proposed method.

Once the target technology is determined, we search for patent documents related to the target technology from patent databases around the world. The collected patent documents are converted into patent keyword data that can be analyzed through the preprocessing process based on text mining. Using the keyword data, we perform the BFA and find factors representing each sub-technology of the target technology domain. In our study, each technological factor is used as an analysis node in the social network analysis. From the results of the social network visualization, we build the technology diagram using degree, closeness, and betweenness. Ultimately, this diagram can be used for various tasks of technology management, such as establishing strategies for technology development, technology forecasting, and new product development.

3.3. Software and Computing Language

The R Project is free software for statistical computing and visualization [30]. It runs on diverse operating systems such as Linux, MacOS, and Windows. In this study, we utilized the R software (version 4.4.0) on a Windows operating system with 16 GB of RAM and an Intel Core i7 processor. Also, we used R data language for statistical computing. R consists of a base module and packages. The base module is included by default when installing R and provides various functions for data analysis and graphics. In contrast, packages are modules that are additionally installed to perform extended functions not provided in the R base. In this paper, we installed and used the ‘tm’, ‘BayesFM’, and ‘sna’ packages for text mining, BFA, and social network visualization, respectively [13,27,31]. To perform the MCMC sampling from the posterior distribution of factor loading, we also used the R base module and packages such as ‘BayesFM’ [27].

4. Experiments and Results

4.1. Simulation to Compare the Performance of FA and BFA

We employed BFA instead of traditional FA to identify latent technologies in the patent keyword data. Unlike FA, which estimates factor loadings using single-point estimates, BFA utilizes probability distributions, providing additional statistical measures such as confidence intervals to account for parameter uncertainty. Specifically, while FA does not incorporate prior information about parameters, BFA introduces a prior distribution, enabling the identification of latent structures even with a small dataset. For this reason,

BFA was used in this study to analyze patent keyword data. In this section, we perform a simulation to compare the performance of FA and BFA. The number of variables and factors for the simulation were set to 6 and 2, respectively. We generated the sample data from the true factor loadings shown in Table 3.

Table 3. True factor loadings.

Variable	Factor 1	Factor 2
X1	0.85	0.05
X2	0.65	0.15
X3	0.75	0.25
X4	0.15	0.75
X5	0.25	0.85
X6	0.05	0.65

To compare the performance differences between FA and BFA, we performed two experiments according to sample size. Figure 5 shows the comparison results of performing FA and BFA using the generated data, with sample size = 500.

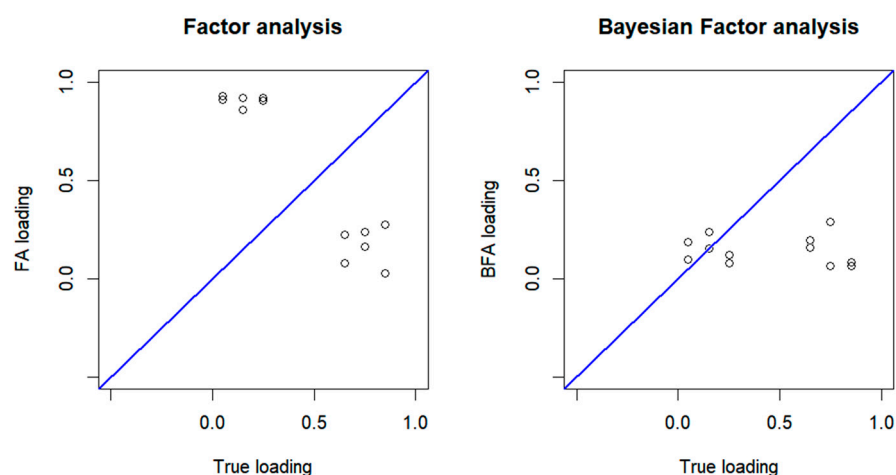


Figure 5. Comparison of FA and BFA with respect to factor loadings; sample size = 500.

The X-axis of the graph represents the true factor loadings, while the Y-axis represents the factor loadings estimated by FA (left) and BFA (right). A greater deviation of the factor loading points from the diagonal indicates poorer model performance, as it fails to accurately reflect the true factor loadings [11,32,33]. The results show that the factor loadings estimated by FA deviated slightly more from the diagonal compared to those estimated by BFA. Subsequently, we increased the sample size to 2000 and compared the FA and BFA performance.

Consistent with the results, Figures 5 and 6 show that the factor loadings estimated by FA deviated further from the diagonal compared to those estimated by BFA. Additionally, we observed that the performance gap between FA and BFA increased as the sample size grew, indicating the superior performance of BFA. Therefore, we found that the performance of the BFA was better than that of the FA. Through the simulation results, we were able to confirm the excellent performance of BFA, once again.

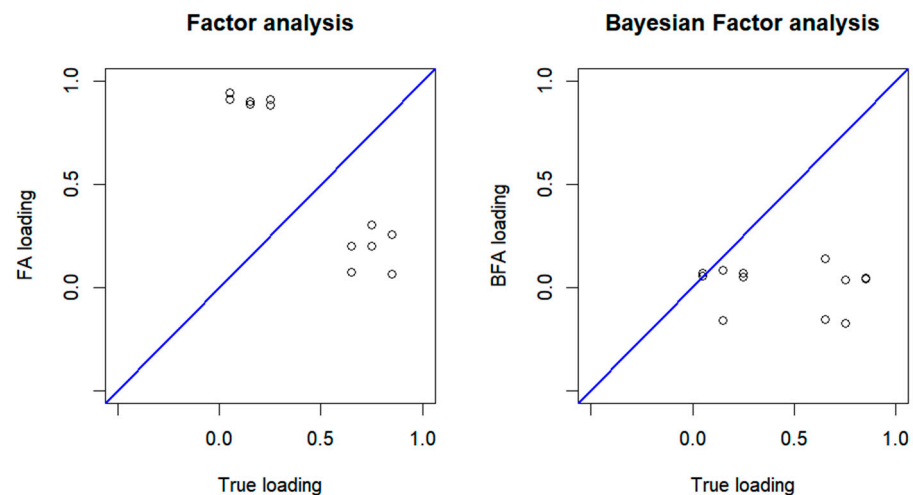


Figure 6. Comparison of FA and BFA with respect to factor loadings; sample size = 2000.

4.2. Experimental Data for Digital Therapy Technology

In this paper, we used the patent documents related to digital remedy and therapeutics. Digital therapy is a method of treating a patient's illness using digital technologies such as software and data [34,35]. We searched the patents from patent databases across the world [36,37]. Using the text mining and valid patent selection processes, we obtained 2685 patent documents and 675 terms. In this experiment, we chose 30 patent keywords highly related to digital therapy from the terms that appeared more than 100 times in all the patent documents. The chosen keywords used in our experiment are as follows: device, data, patient, user, control, information, monitoring, measurement, sensor, therapy, computing, remote, image, interface, signal, display, agent, analysis, network, predict, diagnostics, program, healthcare, brain, electron, machine, database, learn, software, and wireless. Thus, we constructed a matrix consisting of 2685 documents and 30 patent keywords. Each element of this matrix represents the frequency value of a keyword occurring in each patent document. The keywords were considered as variables in our model. Next, we showed the analysis results and their applications in a practical domain.

4.3. Analyzing Patent Keywords of Digital Therapy Technology

Using the 30 patent keywords, we carried out the BFA to understand the technology of digital therapy. First, to determine the optimal number of factors, we calculated the eigenvalues for each factor. Figure 7 shows the eigenvalues for all the factors.

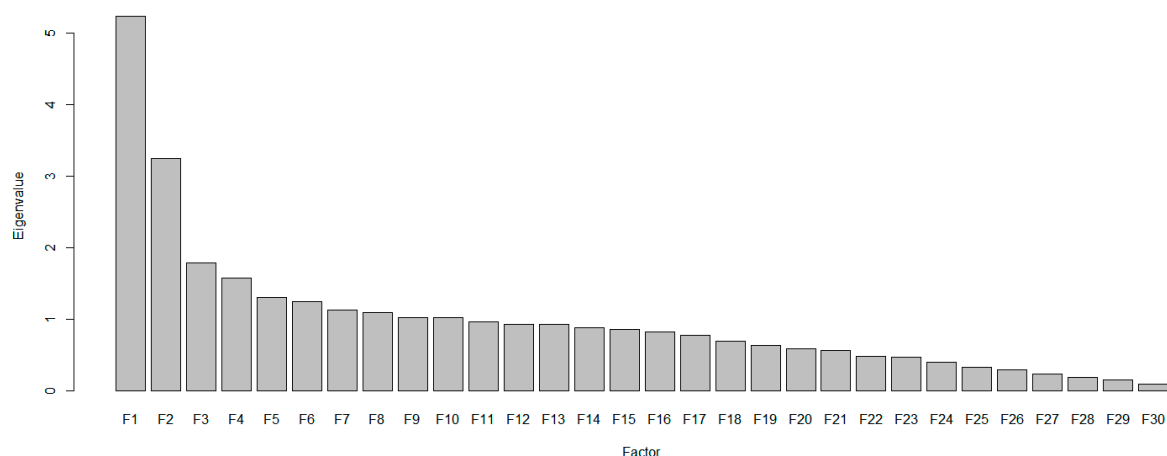


Figure 7. Eigenvalues of all factors.

We can see that among the total 30 factors, the top 2 (F1 and F2) are very large compared to the others. Since the larger the eigenvalue of a factor, the greater the explanatory power of that factor, we present the eigenvalue value and percentile of each factor in Table 4.

Table 4. Eigen value and percentile corresponding to each factor.

Factor	Eigen	Percentile	Factor	Eigen	Percentile	Factor	Eigen	Percentile
1	5.2330	0.1744	11	0.9626	0.0321	21	0.5705	0.0190
2	3.2453	0.1082	12	0.9327	0.0311	22	0.4886	0.0163
3	1.7921	0.0597	13	0.9246	0.0308	23	0.4668	0.0156
4	1.5821	0.0527	14	0.8847	0.0295	24	0.4051	0.0135
5	1.3059	0.0435	15	0.8567	0.0286	25	0.3273	0.0109
6	1.2475	0.0416	16	0.8248	0.0275	26	0.2921	0.0097
7	1.1341	0.0378	17	0.7790	0.0260	27	0.2417	0.0081
8	1.0941	0.0365	18	0.6904	0.0230	28	0.1868	0.0062
9	1.0291	0.0343	19	0.6381	0.0213	29	0.1599	0.0053
10	1.0193	0.0340	20	0.5843	0.0195	30	0.1006	0.0034

As illustrated in Figure 3, the eigenvalues of Factors 1 and 2 were confirmed to be 5.2330 and 3.2453, respectively, which is substantially higher than those of the remaining factors. In accordance with the standard criterion for FA, where factors with eigenvalues equal to or greater than 1 are typically retained, only factors meeting this threshold were included in the analysis. Consequently, the top 10 factors with eigenvalues exceeding 1 were selected for this study. From the results in Table 4, we know that the sum of the explanatory powers of the 10 selected factors is 62.28%. To define each factor as a latent variable, we illustrate the keywords included in each factor and the loading values of the corresponding keywords in Table 5.

Table 5. Keyword list corresponding to each factor.

Factor	Keyword (Factor Loading)
1	device (1.322)
2	data (2.081), database (0.097), analysis (0.088), healthcare (0.088), diagnostics (0.053)
3	patient (1.984), remote (1.079), monitoring (0.811), network (0.527), wireless (0.161)
4	user (1.105), machine (0.156), predict (0.082)
5	control (0.441), electron (0.147)
6	measurement (1.983), computing (0.947), information (0.902), brain (0.767), software (0.710), image (0.261), learn (0.121), agent (0.059)
7	sensor (0.788)
8	interface (0.308), therapy (0.158), program (0.115)
9	signal (0.489)
10	display (0.356)

We confirmed that Factor 1 is a latent variable represented by the keyword device. Next, Factor 2 is represented by five keywords—data, database, analysis, healthcare, and diagnostics—and among these, the keyword ‘data’ has the largest loading value, so we defined Factor 2 centered on the keyword ‘data’. We also defined latent variables for the remaining factors using keywords and loadings, similar to Factors 1 and 2. Therefore, we

performed latent variable definition for all the selected factors and the results are shown in Table 6.

Table 6. Definition of the latent variable corresponding to each factor.

Factor	Technology Definition of Factor
Factor 1	Device system
Factor 2	Data analysis system for healthcare and diagnostics
Factor 3	Remote patient monitoring system using a wireless network
Factor 4	Predictive machine for user state
Factor 5	Electronic control system
Factor 6	Software agent to learn and compute measurement information such as image and brainwave
Factor 7	Sensing system
Factor 8	Patient interface for therapy programs
Factor 9	Signal processing system
Factor 10	Display system

In our experiment, since each keyword corresponds to a detailed technology, the 10 factors defined in Table 6 become representative technologies required for the development of digital remedy and therapeutics technology. Using the results of the BFA, we carried out social network visualizations. First, we show the results of the social network visualization using all the 30 patent keywords, without performing BFA, in Figure 8.

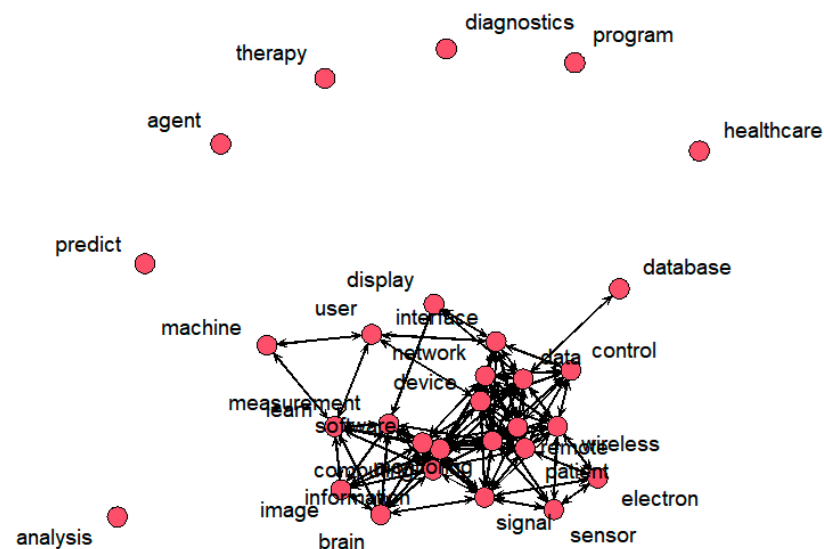


Figure 8. Social network visualization using 30 patent keywords.

In the social network shown in Figure 8, the threshold value of the correlation coefficient for the connections between the keywords was set to 0.15. Various network structures were analyzed by adjusting the threshold, and the final value was selected to construct a social network that best explains digital therapeutics. We found that seven keywords—program, healthcare, agent, analysis, diagnostics, therapy, and predict—despite being patent keywords required for digital therapeutic technology, were isolated and not connected to the entire network group. To gain a more detailed understanding of social network visualization, we performed the visualization again after excluding these seven keywords, and the results are presented in Figure 9.

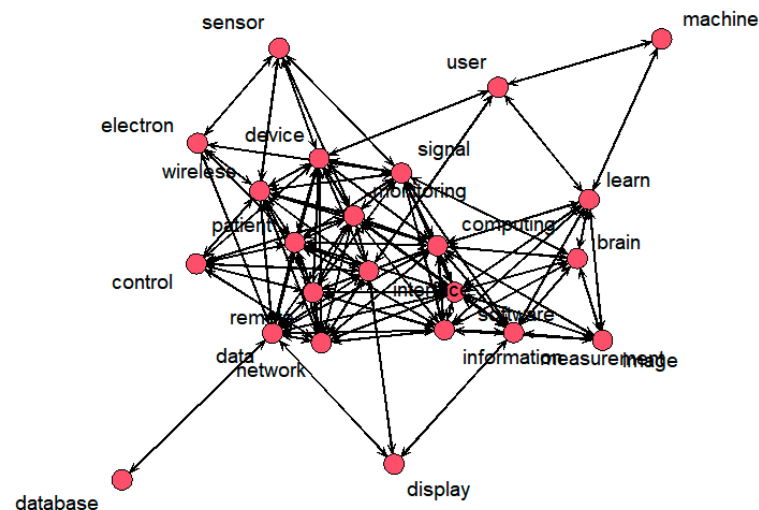


Figure 9. Reduced social network visualization using 23 patent keywords.

Figure 9 is a part of Figure 8. That is, Figure 9 is the result of deleting the seven keywords—analysis, predict, agent, therapy, diagnostics, program, and healthcare—that are isolated and not connected to the network in Figure 8. Therefore, we can better understand the structure of the keyword nodes connected to the network through Figure 9, although, from Figure 8, we can still understand the entire network, including the isolated keyword nodes. It was observed that 12 keywords—device, data, patient, information, monitoring, computing, remote, interface, signal, network, software, and wireless—are positioned at the center of the keyword network, describing digital therapeutic technologies. Additionally, we observed that five keywords—data, computing, signal, learn, and software—function as intermediaries facilitating connections between the other keywords. To confirm the importance and relevance of each keyword, we calculated three performance evaluation measures commonly used in social network analysis (degree, closeness centrality, and betweenness centrality) and show them in Table 7.

Investigating the results in Table 5 from the perspective of degree measure, we can see that they are distributed very diversely, from the keywords ‘monitoring’ and ‘computing’, with the highest degree of 28, to seven isolated keywords with a degree of 0. Also, we find that the closeness centrality is very widely distributed, from the keyword *computing*, which has the highest value of 0.6207, to seven keywords with values of 0. Finally, we can see that the betweenness centrality is widely distributed, from the keyword ‘data’, with the largest value of 62.7908, to 11 keywords with a value of 0. Next, Table 8 shows the top 10 keywords based on degree and centrality.

We found that the results for degree and closeness centrality are similar. That is, we could identify that the top 10 keyword lists in terms of degree and closeness centrality are identical. On the other hand, we confirmed that the results for betweenness centrality are different from those for degree or closeness centrality. From the results in Tables 7 and 8, we can see that the keywords *data*, *information*, *computing*, *signal*, *device*, and *software* are important to developing digital therapy technology. Next, in Figure 10, we present a social network visualization using the top 10 factors and BFA.

From the results in Figure 10, we found that Factors 1, 3, and 9 are important and necessary technologies for digital therapy technology; this is because they are connected to many other factors and are also centrally located in the network. That is, technologies based on device systems, patient monitoring, and signal processing represent the core technologies for digital therapy technology. The technologies of Factors 5 and 7, representing electronic control and sensing systems, respectively, are also central technologies for developing digital therapy technology. We confirmed that data analysis, the most important

technology in digital therapy, is directly connected to four technologies: device system, patient monitoring, software agent, and display system. To investigate the technology network consisting of 10 factors in more detail, we computed the degree and centrality of the factors and show the results in Table 9.

Table 7. Degree and centralities of the 30 keywords.

Keyword	Degree	Closeness	Betweenness
agent	0	0.0000	0.0000
analysis	0	0.0000	0.0000
brain	14	0.4828	1.9152
computing	28	0.6207	35.9638
control	14	0.4713	0.0000
data	26	0.5977	62.7908
database	2	0.3420	0.0000
device	24	0.5862	29.9565
diagnostics	0	0.0000	0.0000
display	6	0.4195	1.9800
electron	10	0.4397	2.0667
healthcare	0	0.0000	0.0000
image	12	0.4540	0.0000
information	24	0.5862	21.6287
interface	20	0.5402	25.0724
learn	16	0.4943	35.0467
machine	4	0.3391	0.0000
measurement	18	0.5287	17.8191
monitoring	28	0.6149	23.8256
network	22	0.5632	5.7263
patient	26	0.5977	14.8674
predict	0	0.0000	0.0000
program	0	0.0000	0.0000
remote	24	0.5805	7.4007
sensor	10	0.4368	1.1500
signal	24	0.5747	25.5908
software	24	0.5862	21.6287
therapy	0	0.0000	0.0000
user	8	0.4368	19.7153
wireless	24	0.5805	13.8555

Table 8. Top 10 keywords by degree and centrality.

Ranking	Degree	Closeness	Betweenness
1	computing	computing	data
2	monitoring	monitoring	computing
3	data	data	learn
4	patient	patient	device
5	device	device	signal
6	signal	information	interface
7	information	software	monitoring
8	software	wireless	information
9	wireless	remote	software
10	remote	signal	user

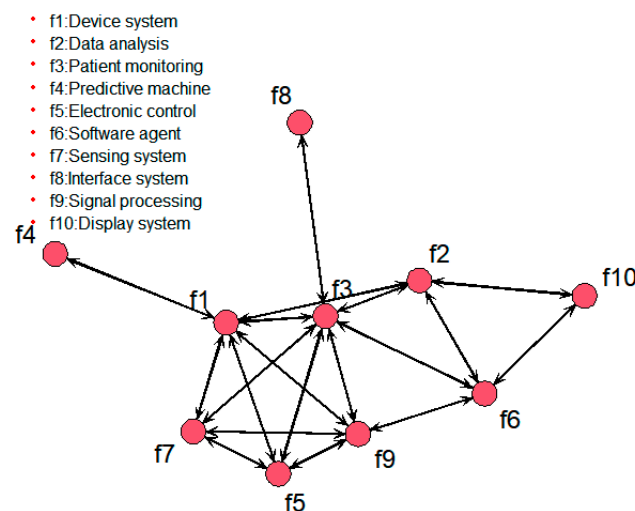


Figure 10. Social network visualization using 10 factors.

Table 9. Degree and centralities of 10 factors.

Factor	Degree	Closeness	Betweenness
factor 1	12	0.8333	19.6667
factor 2	8	0.7222	9.3333
factor 3	14	0.8889	24.0000
factor 4	2	0.5000	0.0000
factor 5	8	0.7037	0.0000
factor 6	8	0.7037	6.6667
factor 7	8	0.7037	0.0000
factor 8	2	0.5185	0.0000
factor 9	10	0.7778	4.3333
factor 10	4	0.5370	0.0000

The degree measure results show that Factors 1, 3, and 9 have values larger than 10. In addition, we found that the closeness centralities of Factors 1, 3, and 9 are ranked in the top three. Thus, these three factors are major technologies in digital therapy technology. On the other hand, we could confirm that the betweenness centrality result is slightly different from those for degree and closeness centrality. From the betweenness centrality result, we can see that Factor 2 is newly included in the top three factors. To compare the importance of all the factors within the network, we rank the factors in terms of degree and centrality in Table 10.

Table 10. Factor ranking by degree and centrality.

Ranking	Degree	Closeness	Betweenness
1	factor 3	factor 3	factor 3
2	factor 1	factor 1	factor 1
3	factor 9	factor 9	factor 2
4	factor 2	factor 2	factor 6
5	factor 5	factor 5	factor 9
6	factor 6	factor 6	factor 5
7	factor 7	factor 7	factor 7
8	factor 10	factor 10	factor 10
9	factor 4	factor 8	factor 8
10	factor 8	factor 4	factor 4

From the results in Table 7, we found that the ranking list of degree and closeness centrality are the same. However, the ranking by betweenness centrality was slightly different from that by degree and closeness centrality. We found that among the top three factors by betweenness centrality, Factor 2 was included instead of Factor 9, which was included in degree and closeness centrality. Using the results in Figure 10 and Tables 6–9, we constructed a technology diagram for digital therapy in Figure 11.

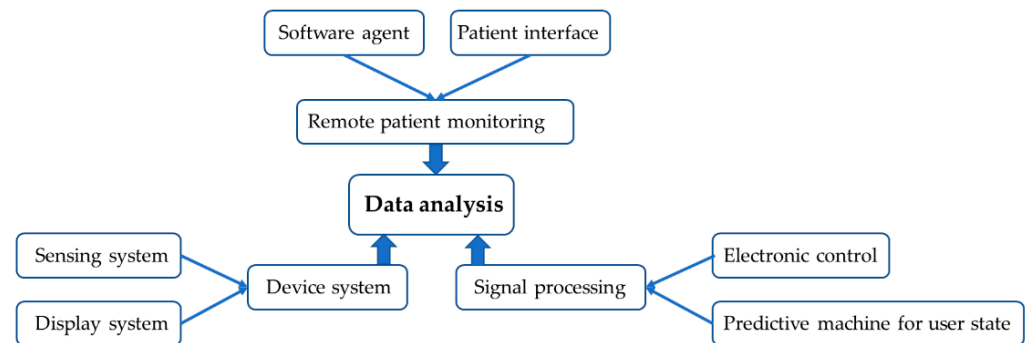


Figure 11. Technology diagram for digital therapy.

We built the technology diagram using the factor definitions described in Table 6 instead of the argument numbers. In Figure 11, the four sub-technologies within the square box represent Factors 1, 2, 3, and 9, while the six sub-technologies outside the box correspond to the remaining factors, excluding the four within the box. We can see that all the sub-technologies required in the field of digital therapeutics are based on data analysis technology. Also, we identified that the technologies related to remote patient monitoring, device systems, and signal processing are central technologies in digital therapy.

5. Discussion

Compared to previous studies on patent keyword analysis, this paper has two major advantages. First, we identified factors that are latent variables in order to find hidden patterns among the detailed technologies of the target technology. To this end, this paper proposes a patent keyword analysis method using BFA, a factor analysis based on Bayesian inference. Second, we constructed a technology diagram for each factor representing detailed technology using social network visualization and closeness and betweenness measures. By obtaining technological factors related to our target technology, digital therapy, and using these to construct a technology diagram, we were able to identify the technological structure between detailed technologies in the digital therapy domain.

From the experimental results, we found that the data technology is most important in digital therapy. Also, the technology of data analysis is influenced by the sub technologies based on remote patient monitoring, device system, and signal processing. That is, it was found that it is most important to process and digitize signals indicating the patient's condition using a device system that can monitor the patient remotely and accurately analyze them to treat the patient. In addition, we were able to confirm that interfaces between patients and computers, software agents, sensing, and predicting patient conditions are also essential technologies in the field of digital therapy.

Our research can be extended from BFA analysis and social network visualization to Bayesian causal inference and Bayesian visualization. We infer technological causal relationships between keywords from a model that applies Bayesian causal inference and use these results to build a Bayesian network [38–41]. Through this, new research will be conducted for an advanced model to identify the causal structure between technology keywords.

6. Conclusions

In this paper, we propose a novel patent keyword analysis method to enhance the understanding of target technology. Traditional patent keyword analysis models have struggled to identify latent technology structures within detailed technologies in a specific domain. To overcome this limitation, we combined BFA with social network visualization to analyze patent keywords. We performed our experiments using patent data related to digital therapy. The patent documents were retrieved from patent databases across the world and preprocessed to construct a patent-keyword matrix. Through BFA, we extracted the factors representing detailed technologies in the digital therapy domain. Subsequently, we employed social network analysis to visualize these factors. Based on the visualization results, we developed a technology diagram to better understand the technological structure of digital therapy.

Our results show that digital therapy technology data analysis is core in the digital therapy domain, and the technologies of remote patient monitoring, device systems, and signal processing are the supporting technologies for data analysis. We also found that technologies based on software agents, patient interfaces, predictive machines for user states, sensing systems, display systems, and electronic controls are necessary for developing digital therapy. This study can contribute to technology management, such as the establishment of technology strategies, analyzing target technologies, and deriving actionable insights across various technological domains in digital therapy technology. This study can be utilized for patent technology keyword analysis for technology management in various domains, as well as in digital therapy technology. In particular, our proposed method is expected to contribute to various technology analyses using patent keyword data.

Our proposed method is mainly based on a statistical approach that applies the Bayesian inference to FA for patent keyword analysis. Using the posterior distribution of Bayesian statistics, we can model uncertainty. In conventional FA, all the parameters, including the factor loadings have fixed constant values; but in BFA, they all follow a probability distribution. Uncertainty cannot be explained by a single constant value, but modeling uncertainty becomes possible by using probability distributions. This is why we propose a model that applies Bayesian posterior probability distribution to FA in this paper. If Bayesian inference is applied to machine learning such as neural networks, we can expect more diverse results and better performance. Thus, in our future work, we aim to apply more advanced statistical approaches to uncover latent structures and further improve the efficiency of understanding target technology. That is, we will study a new patent keyword analysis method that applies Bayesian neural networks, which is Bayesian deep learning, to latent variables obtained by BFA.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to the limitations of the ongoing project.

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

References

1. Roper, A.T.; Cunningham, S.W.; Porter, A.L.; Mason, T.W.; Rossini, F.A.; Banks, J. *Forecasting and Management of Technology*; John Wiley & Sons: Hoboken, NJ, USA, 2011.
2. Hunt, D.; Nguyen, L.; Rodgers, M. *Patent Searching Tools & Techniques*; Wiley: Hoboken, NJ, USA, 2007.
3. Shin, H.; Lee, H.J.; Cho, S. General-use unsupervised keyword extraction model for keyword analysis. *Expert Syst. Appl.* **2023**, *233*, 120889. [[CrossRef](#)]
4. Bzhalava, L.; Kaivo-oja, J.; Hassan, S.S. Digital business foresight: Keyword-based analysis and CorEx topic modeling. *Futures* **2024**, *155*, 103303. [[CrossRef](#)]

5. Jun, S. Keyword Data Analysis Using Generative Models Based on Statistics and Machine Learning Algorithms. *Electronics* **2024**, *13*, 798. [CrossRef]
6. 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]
7. Xue, D.; Shao, Z. Patent text mining based hydrogen energy technology evolution path identification. *Int. J. Hydrogen Energy* **2024**, *49*, 699–710. [CrossRef]
8. Jun, S. Zero-Inflated Text Data Analysis using Generative Adversarial Networks and Statistical Modeling. *Computers* **2023**, *12*, 258. [CrossRef]
9. Jun, S. Patent Keyword Analysis Using Bayesian Zero-Inflated Model and Text Mining. *Stats* **2024**, *7*, 827–841. [CrossRef]
10. Gelman, A.; Carlin, J.B.; Stern, H.S.; Dunson, D.B.; Vehtari, A.; Rubin, D.B. *Bayesian Data Analysis*, 3rd ed.; Chapman & Hall/CRC Press: Boca Raton, FL, USA, 2013.
11. Conti, G.; Frühwirth-Schnatter, S.; Heckman, J.J.; Piatek, R. Bayesian exploratory factor analysis. *J. Econom.* **2014**, *183*, 31–57. [CrossRef] [PubMed]
12. Butts, C.T. Social Network Analysis with sna. *J. Stat. Softw.* **2008**, *24*, 1–51. [CrossRef]
13. Butts, C.T. *Package ‘sna’ Version 2.8, Tools for Social Network Analysis*; CRAN of R Project; R Foundation for Statistical Computing: Vienna, Austria, 2024.
14. Johnson, R.A.; Wichern, D.W. *Applied Multivariate Statistical Analysis*, 6th ed.; Pearson: Essex, UK, 2012.
15. Murphy, K.P. *Machine Learning: A Probabilistic Perspective*; MIT Press: Cambridge, MA, USA, 2012.
16. Theodoridis, S. *Machine Learning: A Bayesian and Optimization Perspective*; Elsevier: London, UK, 2015.
17. Goodrich, M.T.; Tamassia, R.; Goldwasser, M.H. *Data Structures and Algorithms in Python*, 1st ed.; Wiley: Hoboken, NJ, USA, 2013.
18. Sucar, L.E. *Probabilistic Graphical Models Principles and Applications*; Springer: New York, NY, USA, 2015.
19. Feng, L.; Niu, Y.; Liu, Z.; Wang, J.; Zhang, K. Discovering Technology Opportunity by Keyword-Based Patent Analysis: A Hybrid Approach of Morphology Analysis and USIT. *Sustainability* **2020**, *12*, 136. [CrossRef]
20. Hu, M.; Mu, Y.; Jin, H. A bibliometric analysis of advances in CO₂ reduction technology based on patents. *Appl. Energy* **2025**, *382*, 125193. [CrossRef]
21. Yang, X.; Sun, B.; Liu, S. Study of technology communities and dominant technology lock-in in the Internet of Things domain—Based on social network analysis of patent network. *Inf. Process. Manag.* **2025**, *62*, 103959. [CrossRef]
22. Shmueli, G.; Bruce, P.C.; Yahav, I.; Patel, N.R.; Lichtendahl, K.C., Jr. *Data Mining for Business Analytics Concepts Techniques and Applications in R*; John Wiley & Sons: Hoboken, NJ, USA, 2018.
23. Dai, Z.; Zhao, X.; Cui, B. TFIDF Text Keyword Mining Method Based on Hadoop Distributed Platform Under Massive Data. In Proceedings of the 2024 IEEE 2nd International Conference on Image Processing and Computer Applications (ICIPCA), Shenyang, China, 28–30 June 2024; pp. 1844–1848.
24. Hu, H.; Chen, J.; Hu, H. Digital Trade Related Policy Text Classification and Quantification Based on TF-IDF Keyword Algorithm. In Proceedings of the 2024 International Symposium on Intelligent Robotics and Systems (ISOIRS), Changsha, China, 14–16 June 2024; pp. 284–288.
25. Singh, S.; Gupta, S.; Singh, V.; Narmadha, T.; Karthikeyan, K.; Chavan, G.T. Text Mining for Knowledge Discovery and Information Analysis. In Proceedings of the 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 24–28 June 2024.
26. Jain, K.; Srivastava, M. Which Technologies Will Drive the Battery Electric Vehicle Industry?: A Keyword Network Based Roadmapping. In Proceedings of the 2024 Portland International Conference on Management of Engineering and Technology (PICMET), Portland, OR, USA, 4–8 August 2024; pp. 1–9.
27. Piatek, R. *Package ‘BayesFM’ Ver. 0.1.7, Bayesian Inference for Factor Modeling*; CRAN of R Project; R Foundation for Statistical Computing: Vienna, Austria, 2024.
28. Hogg, R.V.; McKean, J.M.; Craig, A.T. *Introduction to Mathematical Statistics*, 8th ed.; Pearson: Upper Saddle River, NJ, USA, 2018.
29. Bruce, P.; Bruce, A.; Gedeck, P. *Practical Statistics for Data Scientists*; O’Reilly Media: Sebastopol, CA, USA, 2020.
30. R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2024. Available online: <http://www.R-project.org> (accessed on 1 April 2024).
31. Feinerer, I.; Hornik, K. *Package ‘tm’ Version 0.7-13, Text Mining Package*; CRAN of R Project; R Foundation for Statistical Computing: Vienna, Austria, 2024.
32. Hansen, B.; Avalos-Pacheco, A.; Russo, M.; De Vito, R. A Variational Bayes Approach to Factor Analysis. In Proceedings of the Bayesian Statistics, New Generations New Approaches (BAYSM 2022), Montréal, QC, Canada, 22–23 June 2023; Volume 435.
33. Samaniego, F.J. *A Comparison of the Bayesian and Frequentist Approaches to Estimation*; Springer: New York, NY, USA, 2010.
34. Nakamura, K.A.; Kim, N. Digital Therapeutics in Hearing Healthcare: Evidence-Based Review. *J. Audiol. Otol.* **2024**, *28*, 159–166.
35. Liu, M.; Schueller, S.M. Integrating Digital Therapeutics with Mental Healthcare Delivery. *J. Health Serv. Psychol.* **2024**, *50*, 77–85. [CrossRef]

36. USPTO, The United States Patent and Trademark Office. Available online: <http://www.uspto.gov> (accessed on 1 April 2024).
37. KIPRIS, Korea Intellectual Property Rights Information Service. Available online: www.kipris.or.kr (accessed on 1 April 2024).
38. Hanif, A.; Ali, S.; Ahmed, A. A Framework for Fault Diagnosis using Continuous Bayesian Network and Causal Inference. In Proceedings of the 2021 IEEE 19th International Conference on Industrial Informatics (INDIN), Palma de Mallorca, Spain, 21–23 July 2021; pp. 1–8.
39. Chen, R.; Lu, Y.; Witherell, P.; Simpson, T.W.; Kumara, S.; Yang, H. Ontology-Driven Learning of Bayesian Network for Causal Inference and Quality Assurance in Additive Manufacturing. *IEEE Robot. Autom. Lett.* **2021**, *6*, 6032–6038. [[CrossRef](#)]
40. Lu, Y.; Zheng, Q.; Quinn, D. Introducing Causal Inference Using Bayesian Networks and do-Calculus. *J. Stat. Data Sci. Educ.* **2023**, *31*, 3–17. [[CrossRef](#)]
41. Ray, K.; van der Vaart, A. Semiparametric Bayesian Causal Inference. *Ann. Stat.* **2020**, *48*, 2999–3020. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.