information

MDPI

Article

Exploring Technology Influencers from Patent Data Using Association Rule Mining and Social Network Analysis

Pranomkorn Ampornphan * and Sutep Tongngam

School of Applied Statistics, National Institute of Development Administration, Bangkapi, Bangkok 10240, Thailand; sutep@as.nida.ac.th

* Correspondence: p.ampornphan@gmail.com; Tel.: +66-2-61547-8993

Received: 15 May 2020; Accepted: 15 June 2020; Published: 22 June 2020

Abstract: A patent is an important document issued by the government to protect inventions or product design. Inventions consist of mechanical structures, production processes, quality improvements of products, and so on. Generally, goods or appliances in everyday life are a result of an invention or product design that has been published in patent documents. A new invention contributes to the standard of living, improves productivity and quality, reduces production costs for industry, or delivers products with higher added value. Patent documents are considered to be excellent sources of knowledge in a particular field of technology, leading to inventions. Technology trend forecasting from patent documents depends on the subjective experience of experts. However, accumulated patent documents consist of a huge amount of text data, making it more difficult for those experts to gain knowledge precisely and promptly. Therefore, technology trend forecasting using objective methods is more feasible. There are many statistical methods applied to patent analysis, for example, technology overview, investment volume, and the technology life cycle. There are also data mining methods by which patent documents can be classified, such as by technical characteristics, to support business decision-making. The main contribution of this study is to apply data mining methods and social network analysis to gain knowledge in emerging technologies and find informative technology trends from patent data. We experimented with our techniques on data retrieved from the European Patent Office (EPO) website. The technique includes K-means clustering, text mining, and association rule mining methods. The patent data analyzed include the International Patent Classification (IPC) code and patent titles. Association rule mining was applied to find associative relationships among patent data, then combined with social network analysis (SNA) to further analyze technology trends. SNA provided metric measurements to explore the most influential technology as well as visualize data in various network layouts. The results showed emerging technology clusters, their meaningful patterns, and a network structure, and suggested information for the development of technologies and inventions.

Keywords: patent analysis; IPC code; patent title; K-means; association rule mining; text mining; social network analysis

1. Introduction

In the development of innovative or creative products, many companies are more likely to carry out research and development (R&D) to determine feasibility and prevent failure before the production and launch of products to a market. Many large and small companies try to establish departments responsible for "strategic invention" and "ideation development", which involve patent analysis activities [1,2].

Patents provide valuable information that companies or organizations use to examine the opportunities and risks of innovative products they are developing, especially finding some approaches to create new products with more inventive steps—so-called "strategic invention". Strategic invention involves using patent data, together with other forms of data analysis, such as marketing surveys, consumer behavior analysis, and assessment of internal capabilities, including business strategies and so on. The process of invention may start from R&D and progress through feasibility assessments of interesting ideas, product design, and product prototype development to industrial production and product release [3–5].

Ideation development can be performed by analyzing the technology domain of interest to obtain an overview of the technology and investment trends for the protection of the technology, including the existence of high-value inventions. The next step is comparing patent data with ideas, new product concepts, or innovations of interest. This process allows us to examine and assess opportunities and competitive risks in the market. Therefore, patent documents have contributed to the development of the industry by disclosing technical contents such as mechanical structure, production processes, quality improvements of products, and so on [3–5].

Patent documents are considered to be an excellent guideline to inventions in a particular field of technology. Patents are protected only in the country of registration. This means that if the patent is not registered in some countries, those countries can use patent documents as a knowledge base in their organizations, for example, R&D institutes, business sectors, and educational institutes. R&D institutes use patents as research guidelines to search for technology gaps and innovations to develop products that meet customers' needs. Business sectors use patents to monitor technology trends, assess their investment capability, and plan to deal with competitors. Educational institutes use patents to find opportunities for technology transfer and to utilize technology's potential to benefit society [5–8].

Analyzing technology trends or domain features of technologies in patent documents depends on the subjective experience of experts. However, accumulated patent documents consist of a huge amount of text data for experts to analyze. Each patent document consists of a front page, detailed specification, claim, declaration, and list of drawings to illustrate the idea of the invention. Usually, it is difficult and time-consuming to process or read the full texts of patents [9]. Therefore, analyzing technology trends using objective methods is more feasible. There are statistical methods applied to patent analysis to identify a technology's overview, investment volume, life cycle, and so on. There are also data mining methods by which patent data can be well classified, such as technical characteristics that can support decision-making in organizations. Analyzing and mining patent data can help to derive information for technology development trend analysis and forecasting [1,2,9–13].

This study focuses on the analysis of technology trends in patents, using patent information from the European Patent Office (EPO). The samples of patent data were from the period 2009–2018, a recent period of technological change [4].

The main objectives of this study were to (1) find existing technology clusters in patent data, (2) find the relationships between associated technologies in each cluster, and (3) explore and visualize the relationships between associated technologies. The proposed data mining methods include K-means clustering, text mining, and association rule mining. The K-means clustering method was first applied to find and group data with natural similarities. Text mining was applied to transform textual data, i.e., patent titles, into a format that could be easily analyzed. Association rule mining was then applied to identify common co-occurrences among data from each cluster. The input data were International Patent Classification (IPC) codes and technical terms (key terms) in patent titles, processed by the proposed data mining methods. The expected output was the technology distribution within each cluster and association rules that identify relationships among patent data. The obtained association rules were later analyzed using Social Network Analysis (SNA). SNA provides different metrics (i.e., degree of centrality, betweenness centrality, closeness centrality, and so on) that can be utilized to explore the density and distribution of technology, allowing us to know the most influential technologies as well as isolated technologies. The association rules that are visualized in the network

patterns can also help us to identify technology trends from the past decade, which we can use as guidelines for developing next-generation technologies.

This paper is organized as follows: Section 2 describes related works that are applied in this study. Section 3 describes the methodologies used in the patent analysis, including data mining methods, text mining, SNA, and our proposed conceptual framework. Section 4 shows the results and analysis of the findings. Section 5 presents concluding remarks on this work.

2. Related Works

2.1. Patent Database

In this study, we investigated patent data from the European Patent Office (EPO) [10]. The EPO's worldwide database, ESPACENET (formerly written as ESP@CNET), contains online data on more than 110 million patent documents from around the world, in various data formats.

The World Intellectual Property Organization (WIPO) [3] defines the International Patent Classification (IPC) code in sections A–H: A: Human Necessities; B: Performing Operations, Transport; C: Chemistry; D: Textiles, Paper; E: Fixed Construction; F: Mechanical Engineering, Lighting, Heating, Weapons; G: Physics; and H: Electricity. The IPC code is an index that is used to classify inventions, using international standards for which technology they belong to and providing a hierarchical system of language-independent symbols for classification of patents and utility models, as shown in Table 1 [3,6].

Table 1. International Patent Classification (IPC) code structure.

Section	Class	Subclass	Main Group	Subgroup
(1st Level)	**(2nd Level)**	**(3rd Level)**	**(4th Level)**	**(5th Level)**
A	43	B	5/00	5/02
Human necessities	Footwear	Characteristic Feature of Footwear	Footwear for Sporting Purposes	Footwear Boots

A patent title is considered to be a useful secondary source of patent data, as shown in Table 2. The WIPO has issued rules for patent titles, which should convey meaning, indicate the subject to which the invention relates, and contain evidence in different categories (product, process, apparatus, use) [3]. The information from titles of inventions provides the development guidelines in a particular form, which is very useful for patent analysis.

Table 2. Examples of patent titles from the European Patent Office (EPO) database.

Patent_ID	IPC_Code	Titles
317553806	C02F	Wave power generator
267832041	C02F	Water reclamation system and method
267805514	C02F	Device and method for automatic wind-power sewage aeration
317599314	C02F	Method and apparatus for water distillation

2.2. Patent Analysis Reviews

Many studies have been conducted on patent analysis to find opportunities in various technology fields. The research related to our study can be summarized as follows:

Kim et al. (2018) [1] proposed a quantitative analysis for patent documents by applying text mining to extracted keywords. The extracted terms or words came from patent documents based on relevant papers, and their authors' keywords. The most representative terms in this study were applied by "frequency-inverse document frequency" or TF-IDF, which can be used to determine the technical characteristics of patent documents. The expected outcome is an increase in the reliability and quality of patent analysis.

Chae and Gim (2019) [2] proposed a model to analyze the technical inventions from patent applications based on IPC (International Patent Classification) and CPC (Cooperative Patent Classification) codes. A "taxonomy tree" will be created using the hierarchical structure of IPC and CPC of each patent, which identifies the invention patterns and technological trends of patent applicants.

Ma et al. (2014) [9] conducted an experiment on Nano-Enabled Drug Delivery (NEDD), using commercial data, the "Derwent Innovation Index" (DII). The patent title and abstract were rewritten by a technical specialist to make the original data clearer. The keywords from the title and abstract were extracted and carefully selected by experts. After that, extracted terms analyzed by specific software tools, "VantagePoint" [program available at www.thevantagepoint.com] and "ClusterSuite" [program developed by J.J. O'Brien, with Stephen J. Carley, at Georgia Tech-to be available at www.VPInstitute.org]. The results had suggested possible innovations and trends for technology in NEDD.

Jun (2012) [10] proposed various data mining methods to forecast technology trends of the Bio-Industry. The data mining methods consisted of three approaches based on "time series analysis", "association rule mining", and "clustering". The results from the "time series analysis" were used to predict the demand of biotechnology, then assign R&D resources of a company to develop biotechnologies. Secondly, the association rule between IPC codes identified key patents to develop or to buy key patents for biotechnology. Lastly, the patent clustering results let us discover vacant areas of biotechnology and detect the disruptive technologies in biotechnology.

Park et al. (2015) [11] proposed a network model to present sustainable technology from patent documents based on the degree of centrality patterns from Social Network Analysis (SNA). The SNA is a network model construction based on graph theory in computer science. The patent document was from the Ford Motor Company [www.uspto.gov]. The IPC codes were used as the elements of vertices. The connection among vertices technologies and sub-technologies suggested the development of new product and services, and R&D planning for future technologies.

Choi et al. (2015) [12] proposed a predictive model to identify the technology transfer in patent information analysis, focusing on the extraction of vacant core technologies and monitoring technological trends. The predictive model applied a social network analysis, linear regression analysis, and decision tree modeling. The construction model was expected to be useful in technology management in commercialization, preventing mismatches from expert opinions and the wasting of R&D resources.

Choi and Song (2018) [13] proposed "a topic modelling-based approach to extract hidden topics from logistic-related patents using Latent Dirichlet Allocation" (LDA). The patenting activity and major assignees of each topic will be investigated. The technology trends from topics were classified as "emerging topic", "declining topic", "dominant topic", or "saturated topic". This helps organizations to understand technological trends, and the general technology landscape in logistics.

Liu et al. (2019) [14] proposed a network theory and social network analysis to investigate the trends of patent collaboration for a smart grid field in China, the so-called "patent collaboration network". The four indicators, i.e., degree centrality, betweenness centrality, closeness centrality, and eigenvector value, were used to identify the positions of technology in a network, such as the influencer (hub), as well as the interconnections, and the importance of technology.

2.3. Summary of Findings and Observations from Related Works

Patent document can be used to analyze technology and innovation trends and to form guidelines to develop new products and services. The results of the patent analysis will be used as decision-making for technology management. Patent data are systematically classified and stored in a database. We can use certain characteristics to discover the hidden patterns in a particular area of technology. The European Patent Office has made the bulk of patent data available for statistical analysis and data mining. The content of non-numerical data, IPC codes, and patent titles can be used to find the answers according to the research objectives. The data mining methods allow us to apply an in-depth

analysis to find new knowledge. There is an existing data mining tool available to process "structured data" (e.g., IPC code) and "unstructured data" (e.g., patent titles). In our study, we were interested in finding relationships from the data mining results and found that SNA is a tool that can visualize the network of relationships as a network graph. This helps us understand the flow of data and their important parts.

3. Methodology

3.1. K-Means Clustering

K-Means clustering is a data mining technique used to group objects or datasets into clusters based on their similarities. The similarity is the total distance between values in each cluster to the centroid, where each centroid has an average cluster value. The closer the distance, the higher the similarity, and vice versa. The measurement of similarity or Euclidean distance can be calculated by

$$d\left(x_j, c_j\right) = \sqrt{\sum_{j=1}^{n}\left(x_j - c_j\right)^2} \quad (1)$$

The grouping of K-means clustering works as follows:

(1) Determine the number of cluster K from the data domain.
(2) Choose K random points from data as centroid.
(3) Set all the data points to the closest cluster centroid.
(4) Recalculate the centroid of newly formed clusters.
(5) Repeat until there is no change in the centroid, i.e., the data points are in their original clusters.

Next, the cluster validation process was applied to find an appropriate number of clusters in patent datasets. One of the cluster validations that is commonly used to compute results from different values of cluster "k" is the average distance between data points and their cluster centroid. The average distance to the centroid, a function of "k", is plotted and the "elbow point" can be used to roughly determine "k" [6,15,16]. From Figure 1, we can see that the value k = 5 is an elbow point since there is a slight bend on both sides of the point.

K	Avg within distance
2	2865.9
3	1605.4
4	1051.1
5	769.44
6	600.68
7	490.88
8	417.71
9	365.09
10	326.5

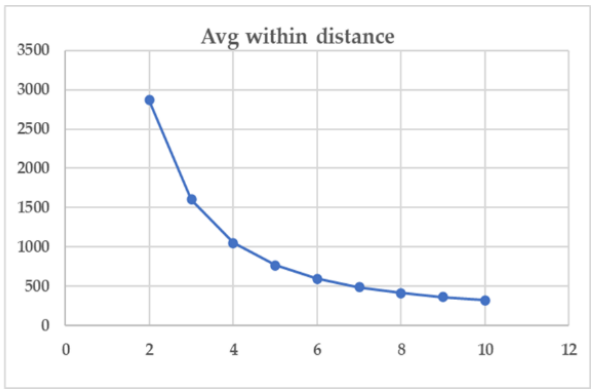


Figure 1. Cluster validation example.

The clustering method is used in market segmentation to find customers that are similar in terms of behaviors. In this study, we applied the marketing approach to determine patent data characteristics. In grouping the patent datasets, we used three attributes (variables)—IPC code, technical fields, technical sectors—to calculate the similarities of each cluster.

3.2. Text Mining

Text mining is a process of knowledge discovery from text documents. The common practice for text mining is the analysis of the information extracted through text processing to form new facts and hypotheses that can be explored further with other data mining algorithms [6,7,15–17].

The major processes of text mining are as follows:

- Tokenizing: the process of breaking text from the document into single words (tokens or terms).
- Filtering out stop words: the process of removing meaningless elements (punctuation marks, special characters, prepositions, articles, pronouns, etc.)
- Transforming cases: the process of transforming all characters into either lowercase or uppercase to avoid confusion between similar words in different cases.
- Stemming: the process of reducing the base form of some single words or their stems.

In the patent analysis, the unstructured text from patent titles of each technology cluster will be preprocessed and transformed into a structured format. The key terms extracted from the text mining approach will be used for further analysis to determine the relationships among invention concepts.

3.3. Association Rule Mining (ARM)

Association rule mining is an algorithm used for discovering interesting relationships between item sets in a large database. The rules from the algorithm can be used to predict existing cases in an item or item set that are grouped. The algorithm uses the parameters support, confidence, and lift to describe the rules that it generates and to select interesting rules from all possible ones. The support is an indication of how frequently the item set appears in the database; the support of rule $(A \rightarrow B)$ can be calculated by the following probability:

$$Support: (A \rightarrow B) = P(A \cup \mathrm{B}). \tag{2}$$

The confidence is an indication of how often the number rule (an if-then statement) is true; the rule $(A \rightarrow B)$ can be represented by conditional probability:

$$Confidence: (A \rightarrow B) = P(B|A). \tag{3}$$

The lift is calculated as the probability of an item set based on the probability of the individual items in the item set; the rule $(A \rightarrow B)$ can be calculated as follows:

$$Lift: (A \rightarrow B) = \frac{P(B|A)}{P(B)} \tag{4}$$

If the rule has a lift greater than 1, it implies that two occurrences are dependent on each other and makes those rules potentially useful for predicting the consequences in future datasets [6,7,15,18].

In patent analysis, IPC codes and the key terms extracted from each technology cluster will be processed to determine the association rules. The association rules (IPC_code #1 → IPC_code #2) determine that "If technology IPC code #1 is developed, then technology IPC code #2 is also developed", and the text association from the extracted technical terms (technical_term #1 → technical_term #2) determines that "If technical_term #1 is developed, then technical_term #2 is also developed". The hidden relationships discovered via association rules help us to summarize the collection of patent documents, in which the IPC code association rules define the technology co-occurrences, and the text association rules derived from the extracted key terms determine the invention concepts.

3.4. Social Network Analysis (SNA)

A social network analysis (SNA) is a study of social connections among actors, such as individuals, groups, organizations, and processes that cause changes in the relationship between individuals,

and between groups, according to the changing situation. SNA helps us understand an informal group, social organization, and the behavior of a social structure. There is a set of measurement metrics to map, measure, explore, and visualize the social relationships between actors. The major metrics include degree centrality, betweenness centrality, closeness centrality, and eigenvector centrality, which are used to analyze and visualize the patterns of network [14,19–22].

The performance metrics used in this study are as follows [22]:

(1) Degree Centrality (DC): a center of connectivity in a network (Hub), which is the most influential in a network. The node that connects many edges is the most influential in a social network. A vertex v of graph G = (V, E) can be calculated as follows:

$$C_D\,(v) = deg\,(v) \tag{5}$$

(2) Betweenness Centrality (BC): the shortest link or path by which an individual node bridges the other node in a network. A high value of BC indicates full control or that it plays an important role between two other nodes participating in a social network. The BC of vertex v of graph G = (V, E) can be calculated as follows:

$$C_B = \sum_{x \neq y \in V} \sigma_{xy}(v)/\sigma_{xy} \tag{6}$$

where σ_{xy} is the total number of shortest paths from node x to node y, and $\sigma_{xy}(v)$ is the number of paths that pass through v.

(3) Closeness Centrality (CC): the mean distance (or average shortest path) from each node to every other node in a network. The high value of CC indicates a broad connection of individuals in a social network. The CC of vertex v of graph G = (V, E) can be calculated as follows:

$$\mathrm{C_C} = \frac{1}{\sum_j d(i,j)} \tag{7}$$

where $d(i, j)$ is the distance between vertex i and j.

(4) Eigenvector Centrality (EC): the relative scores assigned to all nodes in a network. The score of each node is measured from the links with other influential nodes. A high eigenvector score means that a node is connected to many nodes that themselves have high scores. The eigenvector centrality is used for measuring the importance of all nodes in a network. To find the EC score of a graph G = (V, E) with |V| vertices, let B = $b\_\,(v, t)$ be the adjacency matrix, where $b\_\,(v, t) = 1$ if v is linked to vertex t and $b\_\,(v, t) = 0$ otherwise. The relative centrality, x, score of vertex v can be calculated as follows:

$$Xv = \frac{1}{\lambda}\sum\nolimits_{t \in N(v)} x_t = \frac{1}{\lambda}\sum\nolimits_{t \in G} b_{v,t}x_{t^2} \tag{8}$$

where $N(v)$ is a set of neighbors of v, and λ is a constant.

The SNA is the final analysis of this study, by which the IPC code association rules, and text association rules from the technical terms, are visualized as a network graph to determine the technology and invention communities from each cluster. An overview of the network graph lets us see the influential technologies that may have been used to create the invention, and who is the owner of the invention. All of these results can be used as a guideline for technology management to perform R&D and determine the business feasibility.

3.5. Conceptual Framework

The conceptual framework of this study is shown in Figure 2. The patent data, which consist of the IPC code and patent titles, were used as the primary input. The patent data were taken from EPO's online database, then we performed data preprocessing. After that, the four data analysis methods,

K-means clustering, Text Mining, Association rule mining (ARM), and Social network analysis (SNA), were applied to analyze a similar group characteristic of patent data, the hidden knowledge of patent data, and the key influencers of technology and invention. The results are presented in a network graph that identified communities of patent data.

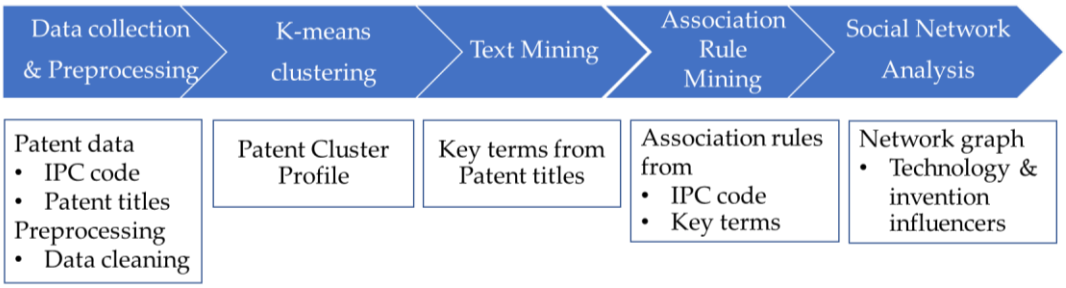


Figure 2. Conceptual framework for patent analysis.

The developed data analysis framework consists of the following steps:

Step 1. Data collection and preprocessing

(1.1) Extract all IPC codes and patent titles from EPO's database.
(1.2) Combine multiple datasets.
(1.3) Perform data cleaning.
(1.4) Transform datasets into a format suitable for K-means clustering and ARM.

Step 2. K-means clustering

(2.1) Perform data clustering to obtain the patent cluster profile.
(2.2) Perform cluster validation to obtain an appropriate number of clusters.

Step 3. Text mining

(3.1) Perform text mining on the patent titles dataset to obtain the technical terms (key terms).

Step 4. Association rule mining (ARM)

(4.1) Apply ARM to IPC code dataset to each cluster to obtain association rules.
(4.2) Apply ARM to technical terms (key terms) to obtain text association rules.

Step 5. Social Network Analysis (SNA)

(5.1) Use SNA to calculate the degree centrality (DC), betweenness centrality (BC), closeness centrality (CC), and eigenvector centrality (EC) of IPC association rules, and the text association rules that exist in each cluster.
(5.2) Construct a network graph to visualize association rules and text association rules in each cluster.
(5.3) Analysis of the results: the most influential technology, connectivity of technology, and technology prioritization, etc.

4. Results and Analysis

4.1. K-Means Clustering

The first analysis process was clustering patent datasets, where the objective was to find existing technology clusters from the patent data. The IPC code, technical field, and technical sector were the

selected variables for the cluster validation process since they were important parts of patent data to identify cluster characteristics. The results of the five clusters show a group of patents, including from chemistry, electrical engineering, instrument, mechanical engineering, and other fields. There were 153,071 patents from 2009–2018 distributed in each cluster and calculated as a percentage, shown in Table 3.

Table 3. K-Means clustering results (K = 5).

Cluster	Technical Sector	Number of Patents	Percent (%)
1	Chemistry	7909	4.84
2	Electrical Engineering	36,322	22.26
3	Instruments	11,570	7.09
4	Mechanical Engineering	97,270	59.63
5	Other Fields	10,046	6.15

The clustering provided the largest number of patents and three clusters with a relatively small number of patents. Cluster 4 has the largest number of patents. This cluster represents an adequate technical sector since it has a large number of patents registered. On the other hand, clusters 1, 3, and 5 are inadequate technology clusters since they have a small number of patents. Each cluster contains data that reflect the relationship between the IPC codes and the key terms extracted from patent titles. Both can be used to describe the technologies, inventions, and influencers that are useful for R&D and technology management in the future. The analysis of IPC codes and the key terms will be explained in the next sections.

Patent cluster characteristics, based on the technical sector, technical field, and IPC code, are summarized in Table 4. Each cluster consists of specific technical fields and IPC codes. The IPC codes represent the inventions shown in each cluster, and will be used to find the relationships between technology by applying association rule mining to forecast technology trends.

Table 4. Results of patent clustering.

Cluster	Technical Sector	Technical Fields	IPC Codes
1	Chemistry	Organic Fine Chemistry	A61Q, C07B, C07C, C07D, C07F
		Biotechnology	C07G, C07K, C12M, C12N, C12P
		Pharmaceuticals	A61P
		Macromolecular Chemistry, Polymer	C08B, C08C, C08F, C08G, C08H
		Food Chemistry	A01H, A21D, A23B, C12C, C12G
		Basic Materials Chemistry	A01N, A01P, C05B, C05C, C05D
		Materials, Metallurgy	B22C, B22D, B22F, C01B, C01C
		Surface Technology, Coating	B05C, B05D, B32B, C23C, C23D
		Microstructural, Nanotechnology	B81B, B81C, B82B, B82Y
		Chemical Engineering	B08B, C14C, D06B, F25J, H05H
		Environmental Technology	A62C, B09B, C02F, F01N, G01T
2	Electrical Engineering	Electrical machinery, Apparatus, Energy	F21K, F21L, G06C, H01B, H01C
		Audiovisual Technology	G09F, G09G, G11B, H04R, H04S
		Telecommunications	G08C, H01P, H01Q, H04B, H04H
		Digital Communication	H04L, H04W
		Basic Communication Processes	H03B, H03C, H03D, H03F, H03G
		Computer Technology	G06C, G06E, G06F, G06G, G06K
		IT Methods for Management	G06Q
		Semiconductors	H01L
3	Instruments	Optics	G02B, G02C, G02F, G03B, H01S
		Measurement	G01B, G01C, G01D, G01F, G01G
		Control	G05B, G05D, G05F, G07B, G07C
		Medical Technology	A61L, A61M, A61N, G16H, H05G

Table 4. *Cont.*

Cluster	Technical Sector	Technical Fields	IPC Codes
4	Mechanical Engineering	Handling	B25J, B65B, B65C, B65D, B65G
		Machine Tools	A62B, B21B, B21C, B21D, B21F
		Engines, Pumps, Turbines	F01B, F20C, F03D, G21B, G21C
		Textile, Paper Machines	A41H, A43D, B41M, C14D, D01B
		Other special Machines	A01B, A01C, B28C, C03B, F41A
		Thermal Processes and Apparatus	F22B, F22D, F22G, F23B, F23C
		Mechanical Elements	F15B, F15C, F15D, F16B, F16C
		Transport	B60B, B60C, B60D, B60F, B60G
5	Other Fields	Furniture, Games	A47B, A47C, A47F, A47G, A47H
		Other Consumer Goods	A99Z, B42D, D04D, F25D, G10B
		Civil Engineering	E01B, E01C, E01D, E01H, E02B

4.2. Text Mining

The K-means clustering performed in Section 4.1 provided the results of five technical sectors: Chemistry, Electrical engineering, Instruments, Mechanical engineering, and Other fields (patents that cannot be identified with any sector). Next, we performed the major processes, i.e., tokenizing, filtering out stop words, transforming cases, and stemming, for extracting the key terms from patent titles. The examples of patent titles and the extracted key terms are shown in Tables 5 and 6, respectively.

Table 5. Examples of patent titles from the technical sector.

Cluster	Technical Sector	Examples of Patent Titles
1	Chemistry	- Method for manufacturing resin impregnated multi-orientation composite material. - Hydrogen supplementation fuel apparatus and method. - Resin transfer molding process for an article containing a protective member.
2	Electrical Engineering	- Power storage and power transfer method and apparatus. - Active power optimizing and distributing method for wind generator unit of wind power station. - Street lamp with power supply system powered by wind heat energy.
3	Instruments	- Wind turbine blade load sensor. - Apparatus and method for automatically fabricating tape with threads for visualization of air streams on aerodynamic surfaces. - Method for sensing strain in a component in a wind turbine, optical strain sensing system and uses thereof.
4	Mechanical Engineering	- Wind turbine comprising a thermal management system. - Electrical power generation via the movement of a fluid body. - Integrated control apparatus and method for hybrid type wind turbine system.
5	Other Fields	- Tower for a wind farm with flange piece for connection of segments. - Waste-receiving device for incontinent persons. - Hydraulic geofracture energy storage system.

The patent titles are considered to be "unstructured text", usually analyzed by experts—different from analyzing variables (IPC codes in this case) that are computer-readable. Applying text mining, the patent titles in each cluster are broken down into smaller units and structured to make the extracted key terms more meaningful.

The key technical terms derived from the text mining of each cluster were the most frequent words found in patent titles. The inventions might have special qualities that were initially defined by the definition of the patent titles. For example, the terms "system", "device", and "process" were commonly found in all clusters. This was because the patents came from the ideas of systems, devices, and processes initiated by the inventors or experts in each technology area. At the same time, many words convey the meaning of inventions that are relevant to each cluster characteristic as well.

Table 6. Examples of extracted key terms from patent titles.

Cluster	Technical Sector	Examples of Extracted Technical Terms
1	Chemistry	system, composite, wind, generator, device, energy, power, turbine, material, blade, coat, structure, water, manufacture, process, product, compound, product, apparatus
2	Electrical Engineering	generator, wind, device, electric, control, energy, turbine, apparatus, machine, solar, magnet, operator, motor, supply, use, converter, base, plant, grid, storage
3	Instruments	wind, device, power, control, turbine, generator, monitor, apparatus, detector, measurement, testing, energy, sensor, electric, blade, base, usage, operator, determinator
4	Mechanical Engineering	apparatus, base, control, converter, device, electric, energy, generator, grid, machine, magnet, motor, operator, plant, solar, storage, supply, turbine, use, wind
5	Other Fields	wind, tower, system, turbine, structure, power, foundation, device, generator, energy, installer, construct, support, assemble, apparatus, concrete, plant, water, type

4.3. Association Rule Mining (ARM)

In the five clusters from clustering results, we used ARM to find the relationship between IPC codes as well as the relationship between key terms. The association rules were applied to find the relationship between the IPC codes that determined antecedent (IPC code #1) and consequent (IPC code #2) of technologies within each cluster. The key terms derived from the text mining process were so-called "text association rules". The text association rules from the key terms determined the relationship between the related terms of invention. Tables 7 and 8 show some examples of association rules between IPC codes as well as some examples of association rules between key terms, with at least a 10% confidence value.

Table 7. Examples of association rules between IPC codes and their support, confidence, and lift.

Cluster	Antecedent	Consequent	Support (%)	Confidence (%)	Lift
Chemistry	C22C	C21D	2.2	51.3	16.3
	C08L	C08K	2.5	49.1	12
	C01B	B01J	1.8	33.6	8.6
	C08G	C08L	1.8	33.4	6.6
	C08G	C08K	1.1	20.9	5.1
Electrical Engineering	F21V	F21S	1.3	52.1	23.5
	G06F	G06Q	1.2	13.4	2.8
	H02M	H02J	2.7	44.6	1.7
	H01M	H02J	1.2	37.2	1.4
	H02M	H02P	1.1	18.3	1.3
Instruments	A61N	A61B	0.5	10.3	12
	A61N	A61M	0.3	39.4	11.3
	G01K	G01W	0.3	17.1	4.9
	A61F	A61M	0.3	17.1	4.8
	A61B	A61M	0.6	14	4
Mechanical Engineering	F01D	F02C	1.3	22.2	8.1
	B63B	F03D	1.2	44	7.1
	F16H	F03D	1.6	39.4	6.3
	F03B	F03D	3.5	38	6.1
	B29C	F03D	1.3	35.7	5.7
Other Fields	E04B	E04C	1.3	16.5	3.8
	E04C	E04H	2.3	54.5	1.6
	E04H	E04B	3.9	11.6	1.4
	E04G	E04H	2	37.4	1.1
	E02D	E02B	3.4	14.7	1

Table 8. Examples of association rules between the key terms and their support, confidence, and lift.

Cluster	Antecedent	Consequent	Support (%)	Confidence (%)	Lift
Chemistry	fiber	reinforce	1.4	40.4	11.7
	wind, turbine	blade	3	59.9	9.4
	wind, generator	power	2.2	55.2	6.3
	system, power	generator	1.6	53.6	5.7
	fiber	material	1.1	31.5	4.2
	structure	composite	2.6	41.8	2.7
	hydrogen	system	1.6	49.8	2.6
	composite	material	2.8	18.4	2.4
Electrical Engineering	storage	energy	2.1	70.2	5.5
	electric	machine	3.3	22.6	4.4
	machine	electric	3.3	66.1	4.4
	wind, control	turbine	1.8	36.8	3.6
	plant	wind	1.8	60.5	2.3
	wind, device	generator	2.2	62.7	2.3
	generator, solar	wind	1.1	58.1	2.2
	wind, solar	generator	1.1	55.3	2
	magnet	generator	1.5	36.4	1.3
Instruments	wind, blade	turbine	2.5	75.1	5.9
	wind, power	generator	3	53.4	4.4
	wind, turbine	blade	2.5	23.6	4.4
	wind, test	turbine	1.3	51.9	4
	wind, device	generator	1.9	47.2	3.9
	turbine, monitor	wind	1.4	85.8	3.8
	turbine	wind	10.7	84	3.7
	device, power	generator	1.3	43	3.6
Mechanical Engineering	method, rotor	blade	1.1	66.8	5.2
	method, blade	rotor	1.1	31.3	5.1
	turbine, rotor	blade	2	62.8	4.9
	turbine, blade	rotor	2	27.7	4.5
	rotor	blade	3.4	55.5	4.3
	plant	power	2.8	78.2	3.3
	method, power	control	1.2	25	3.1
	driven	generator	2.8	80.7	3
	power	generator	14.7	63.9	2.4
Other Fields	wind, power	plant	1.6	22.4	5.9
	wind, structure	support	1.2	33.7	5.9
	wind, plant	power	1.6	60.8	4.5
	turbine, foundation	wind	3.1	98.9	3.6
	turbine, installer	wind	1.7	97.5	3.6
	concrete	tower	2.4	52.8	2.6
	plant	wind	2.7	71.9	2.6
	plant	tower	1.3	34.1	1.7
	structure, support	wind	1.2	44.1	1.6
	wind, system	tower	1	31.9	1.6
	tower, concrete	wind	1	44	1.6

Each cluster identified the top association rules of developed technologies and inventions. The association rules implied that if technology IPC #1 was developed, technology IPC #2 was also developed. Additionally, the output obtained from Section 4.2 was applied ARM to extract text association rules to identify the relationships between key terms in each cluster. There are three common measures to describe association. The results in Tables 7 and 8 can be discussed as follows:

1. The rules with high support value implied the popularity of technologies and inventions. For example, the rule (C08L → C08K) and the rule (wind, turbine → blade) from the Chemistry

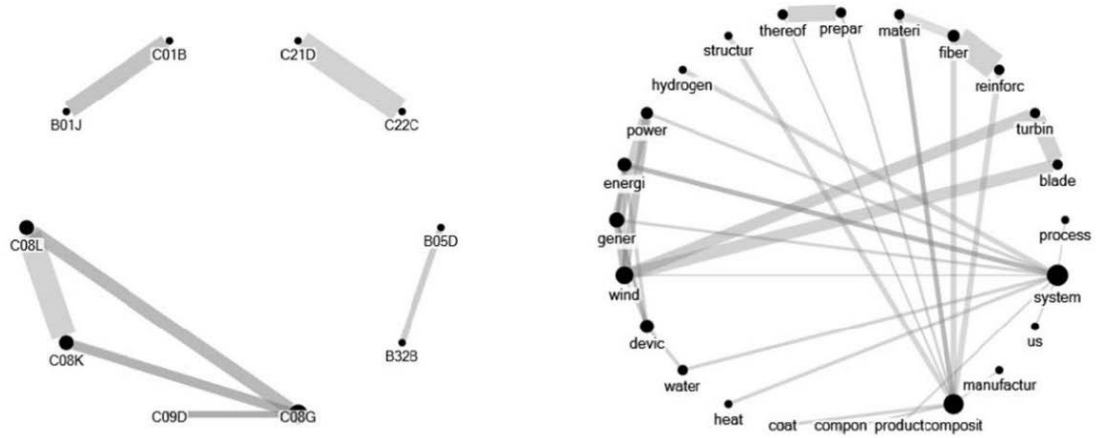
cluster had the highest support value. This means technology C08K was widely developed on technology C08L and the invention of "blade" was widely developed from the invention of "wind" and "turbine".

2. The rules with high confidence value implied the probability of technologies and inventions. For example, the rule (E04C → E04H) and the rule (turbine, foundation → wind) in the Other Fields cluster had the highest confidence values. If technology "E04C" was developed, then technology "E04C" was more likely to developed as well. Additionally, if the inventions related to "turbine" and "foundation" were developed, the invention related to "wind" was more likely to be developed.
3. The rules with high lift value implied a strong relationship between technologies and inventions development. The lift values from each cluster were greater than 1, which means the antecedent and the consequent of the technologies and inventions are more likely to associate with each other. The rules in the first order of each cluster had the highest lift, which means the technologies, as well as the inventions, are dependent on each other and the rules are potentially useful to predict the consequences in the future.

4.4. Social Network Analysis (SNA)

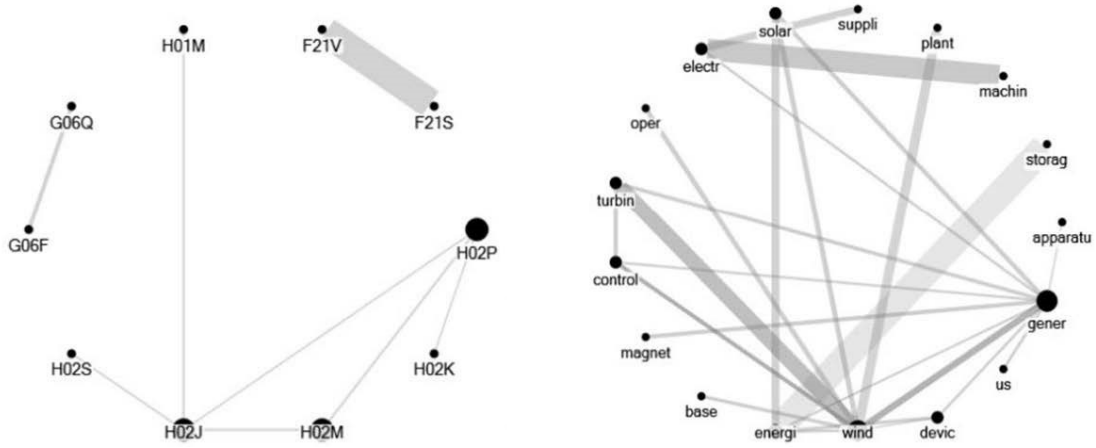
4.4.1. Constructing a Network of ARM

The association rules in each cluster that resulted from Section 4.3 were jointly analyzed by SNA, where the number of rules in each cluster must be large enough to illustrate a network. In this subsection, we used as many rules as possible to illustrate the unambiguous network. The network graph was arranged in a circular layout. Both the antecedent and consequent from the association rule were represented as IPC code and key term vertices in SNA. The size of each vertex depends on the value of the degree centrality. The higher the degree centrality, the greater the vertex size. The lift values from the association rules were represented as the edges that connected the vertices in SNA. The size of each edge depended on the value of the lift. The higher the lift, the greater the scale of the edge size. Figures 3–7 illustrate the relationships among IPC codes and key terms in the Chemistry, Electrical Engineering, Mechanical Engineering, Instrument, and Other Fields, respectively.



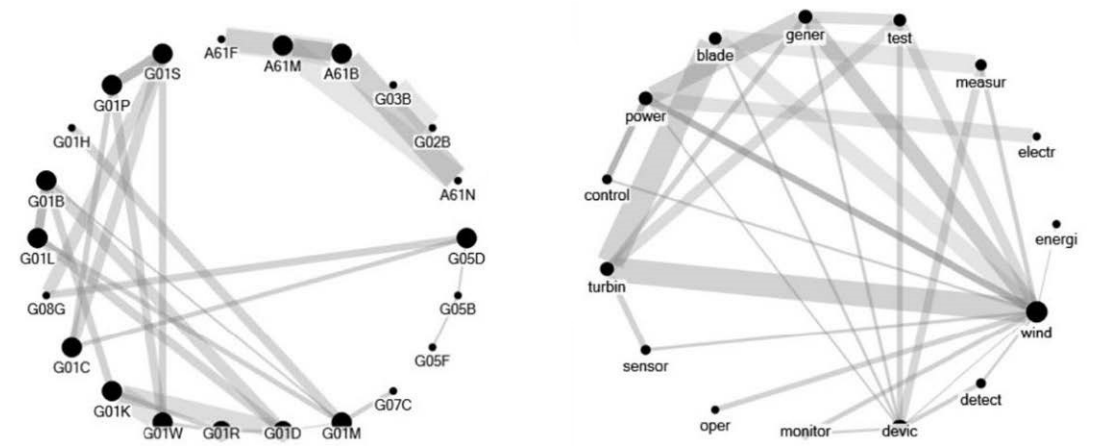
(a) (b)

Figure 3. Chemistry network graph to visualize Association Rule Mining (ARM). (**a**) Chemistry IPC Network; (**b**) Chemistry Key Terms Network.



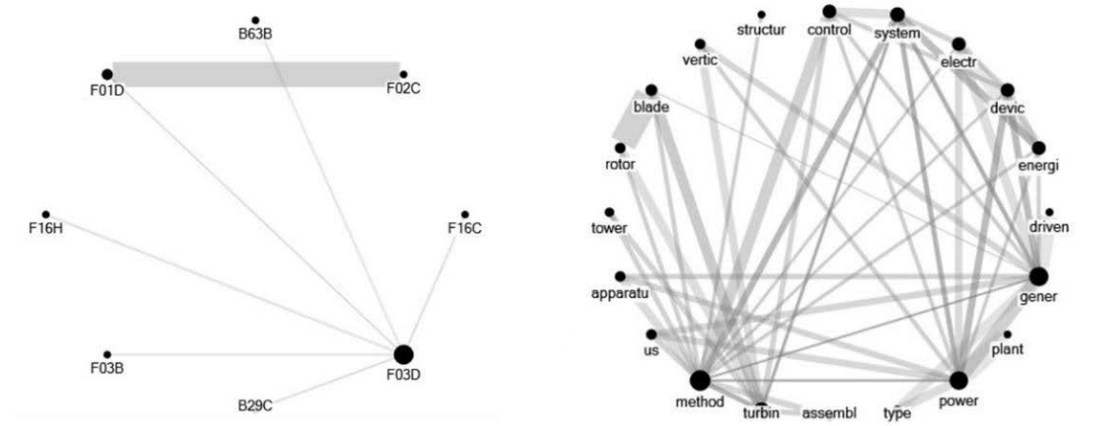
(a) (b)

Figure 4. Electrical engineering network graph to visualize ARM. (**a**) Electrical engineering IPC Network; (**b**) Electrical engineering Key Terms Network.



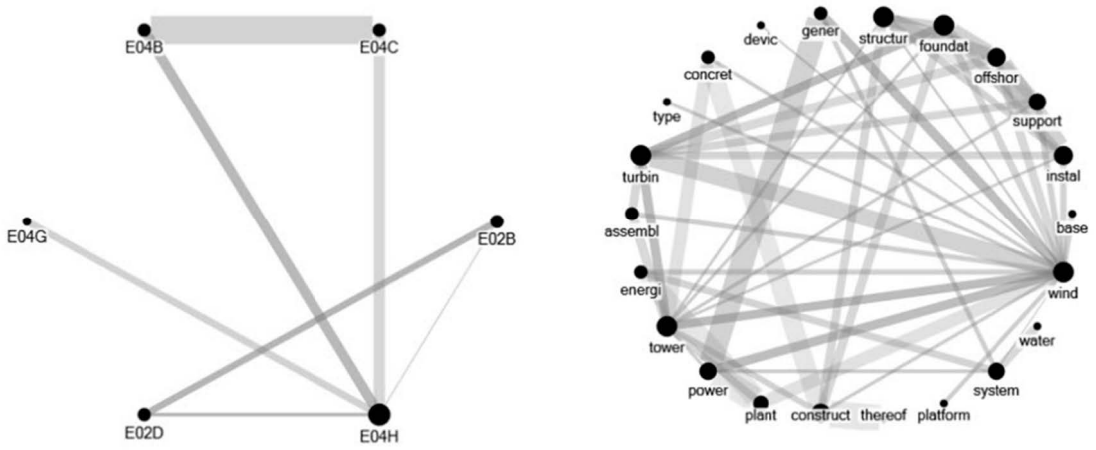
(a) (b)

Figure 5. Instruments network graph to visualize ARM. (**a**) Instruments IPC Network; (**b**) Instruments Key Terms Network.



(a) (b)

Figure 6. Mechanical engineering network graph to visualize ARM. (**a**) Mechanical engineering IPC Network; (**b**) Mechanical engineering Key Terms Network.



(a) (b)

Figure 7. Other fields network graph to visualize ARM. (**a**) Other fields IPC Network; (**b**) Other fields Key Terms Network.

The network graph above shows the IPC codes and key terms in each cluster that represent technology and invention influencers. The most popular technology and invention can be seen from the size of the vertices. The size of each edge determines the possible inspiration of the inventions. For example, technologies C08L, C08K, and C08G are popular (influencers) in the Chemistry cluster. Technologies B01J, C01B, C21D, and C22C are less popular, but they are still inspiring. Therefore, the network graph allows us to visually evaluate the properties of the large number of association rules.

4.4.2. Summary of Influential Nodes from SNA

The network represented the relationships between five clusters with IPC codes and the key terms by degree centrality (DC), betweenness centrality (BC), closeness centrality (CC), and eigenvector centrality (EC). The top ranks of important nodes in the network graph from the results can be summarized as shown in Table 9.

The degree of centrality indicates the hub nodes in the network, which reflects the most influential technology and invention in each technical sector. It is seen that if the nodes are always the first place for all measurements, they control the network. Additionally, they collaborate with other nodes and play important roles in promoting new technology and invention in their sectors. From the degree centrality, the technologies "C08G", "H02J", "G01D", "F03D", and "E04H" are considered to be the most influential technologies, while the key terms related to "system", "generator", "wind", and "method" are the most influential inventions. From the betweenness centrality, these technologies and inventions seemed to cooperate with others in the network. This leads to knowledge exchange. From the closeness centrality, these technologies and inventions are potentially used to develop new products and services. However, there are some isolated technologies in the network, such as in the Chemistry, Electrical engineering, and Instrument sectors, with high CC values. We can assume that these are developed for a specific purpose and are unreachable by the other technologies. We observed that all the key terms in the Other fields sector have equal value. This means these are general key terms of invention and they can be used together. Although there is an invention related to "wind" in the Instruments and Other fields sectors, the developed technologies are different. The Instruments sector is involved with the measurement technology, while the Other fields sector is related to construction in Civil engineering. Lastly, the eigenvector centrality of the nodes "C08G", "H02J", "G01D", "F03D", "F01D", and "E04H" determines the most important technologies to the other nodes from each technical sector, while the most important inventions to the other nodes are related to "system", "generator", "wind", and "method".

Table 9. Results of Social Network Analysis (SNA).

Technology Sector	Influential Nodes	Degree Centrality	Betweenness Centrality	Closeness Centrality	Eigenvector Centrality
Chemistry	IPC code	C08G	C08G	B05D, B32B B01J, C01B C21D, C22C	B32B
	Key terms	system	system	system	system
Electrical Engineering	IPC code	H02J	H02J	G06F, G06Q F21V, F21S	H02J
	Key terms	generator	generator	generator	Generator
Instruments	IPC code	G01D	G01W	A61M, A61B	G01D
	Key terms	wind	wind	wind	Wind
Mechanical Engineering	IPC code	F03D	F03D	F03D	F03D, F01D
	Key terms	method	turbine	method generator power	Method
Other Fields	IPC code	E04H	E04H	E04H	E04H
	Key terms	wind	wind	wind, turbine system, generator, power, device, method, tower, energy, composite	Wind

4.4.3. Application of the Results to Patent Management

The SNA results in the previous subsection not only help decision-makers to evaluate information based on visualization, but also provide measurements (as mentioned above) to determine the connectivity characteristics. The technology influencers of each technical sector and their definitions are shown in Table 10.

Table 10. Technology influencers.

Technical Sector	Technology Influencer	Defined Technology
Chemistry	C08G	Macromolecular chemistry, polymers; Reaction involving carbon to carbon.
Electrical Engineering	H02J	Electrical machinery, apparatus, energy; Circuit arrangement, system for supplying and storing electric power.
Instruments	G01D	Measurement; measuring apparatus for two or more variable.
Mechanical Engineering	F03D	Engines, pumps, turbines; Machines or engines for liquids.
Other Fields	E04H	Civil engineering; Buildings or like structures for particular purposes

When the inventor or company is interested in creating or developing products that are classified in various technical clusters, they will have to check whether other investors hold patents to prevent intellectual property infringement. The technology influencers have been patented by inventors and companies around the world. The number of patents implies the capabilities of technology development in each country. The top five countries of technology influencers, according to the number of patents of each cluster, is shown in Figure 8. The USA and Germany are the countries that have

an impact on technology development in Chemistry, Electrical Engineering, Instruments, and Other Fields, while Korea has an impact on technology development in Mechanical Engineering.

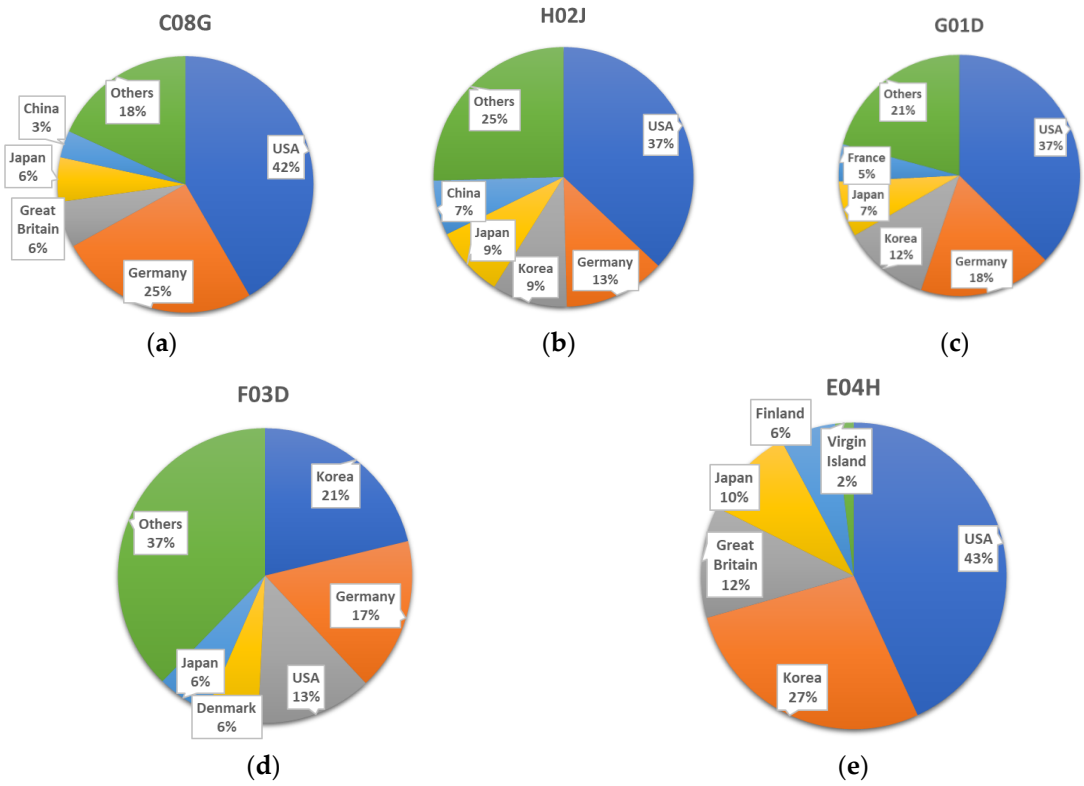


Figure 8. Top five countries in terms of numbers of patents registered. (**a**) Chemistry cluster; (**b**) Electrical engineering cluster; (**c**) Instrument cluster; (**d**) Mechanical engineering cluster; (**e**) Other Fields cluster.

Patent management in an organization is not only about inspecting the patent documents registered by competitors, but also obtaining the appropriate technology to develop products or services. Patent data have inspired the inventor or company to be more creative in producing and upgrading products or services. Product development may be blocked by the inventor or company who holds the patent related to the particular technology influencers. Many companies seek partnership for technology transfer as well as to explore the patents that have not been renewed for "freedom to operate". Although the patents have no novelty, the core technology can be used to further develop new products, process, and services for customers, and this does not infringe on the intellectual property of others.

5. Conclusions

This paper proposes technology analysis from patent documents using IPC codes and patent titles to identify hidden information. The patent data were collected from the European Patent Office (EPO). Our study applied a conceptual framework to find existing technology clusters from the collected patent data, then find the relationships of associated technologies in each cluster, and explore and visualize the insight relationships of associated technologies. The design framework consisted of data mining methods and Social Network Analysis (SNA), which can be useful for the development of new technology and inventions.

The data mining methods, K-Means clustering, Association Rule Mining, and Text mining, were used to analyzing patent data. The K-Means clustering was applied to find the group similarities of patent data to find existing technology clusters from patent data. By performing cluster validation to find an appropriate number of clusters, we observed five clusters that represented technology clusters, i.e., Chemistry, Electrical engineering, Instruments, Mechanical engineering, and Other

fields. The knowledge gained from K-Means clustering was the adequate technology, i.e., Mechanical engineering, since it had the largest amount of patent data. The most inadequate technology was Chemistry, since it had the smallest amount of patent data. Both have some interesting aspects to be developed in the future in order for companies to gain a competitive advantage.

The five technology clusters were the focus groups, where each group consisted of various IPC codes and patent titles. Useful information can be extracted using Association Rule Mining (ARM) and Text Mining. ARM was applied to find the co-occurrence among IPC codes and patent titles. The antecedent (A) and consequent (B) of association rules were defined as; if technology A was developed, then technology B was also developed. ARM helped us deduce meaningful rules that identify important relationships among technology classes and invention concepts.

Text Mining was applied to extract key terms from patent titles. Key terms were extracted from the patent titles in each cluster based on text mining methods. The limitation of this study was that we only considered patent titles in English. The key terms extracted from each cluster were pruned to obtain the most relevant and were counted and indexed to compute the total term occurrence and frequency. The extracted key terms would be used to find the co-occurrences among invention concepts by ARM.

Association rules derived from IPC codes and key terms of patent titles can be assessed by using the values of support, confidence, and lift to determine the strength of the rules. Additionally, we used SNA to further analyze association rules and to visualize a network structure. SNA provided a network visualization and some measurements, i.e., degree centrality, betweenness centrality, closeness centrality, and eigenvector centrality. These factors can be used to determine the most influential technology, as well as the most influential inventions. Additionally, some factors determine the bridges, closeness, and level of importance of technologies and inventions.

The results of the proposed methods and conceptual framework show the relationships in patent data. Each technology cluster consists of the most influential technologies and inventions, connected with each other, and there are opportunities for the development of new technologies and inventions. The technology influencers can be inspired by an inventor or company to develop products or services that satisfy their customers. Many companies search for patents to explore the target technology to develop their products or services as well as avoid intellectual property infringement. Patent management is necessary for companies that require R&D to create new technology for product development. The companies can manage their knowledge by accessing the patents held by individuals or organizations. Access to technological knowledge can be achieved through collaborations between patent holders in order to receive technology transfer. One of the good practices to minimize the risk of infringement on the patent right of others and save companies' resources is to apply for "freedom to operate" during an early stage of the company's establishment.

Summarizing the above, in this study, we applied various data mining methods to gain insight from patent data, and applied SNA to explore technology-influenced networks and investigate the influential patent holders in various technology sectors around the world. This will contribute to making the strategic invention of inventors or companies more effective.

Author Contributions: P.A. designed and performed the experiment, and analyzed the data. S.T. supervised the research and revised the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Institute Development Administration (NIDA), Thailand.

Acknowledgments: Our thanks to the European Patent Office (EPO) for allowing us to access the patent database and update our information regularly.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kim, J.; Choi, J.; Park, S.; Jang, D. Patent keyword extraction for sustainable technology management. *Sustainablility* **2018**, *10*, 1287. [CrossRef]
2. Chae, S.; Gim, J. A study on trend analysis of applicants based on patent classification systems. *Information* **2019**, *10*, 364. [CrossRef]
3. WIPO. World Intellectual Property Organization. 2018. Available online: https://www.wipo.int/classifications/ipc/en/ (accessed on 1 August 2019).
4. EPO. European Patent Office, ESPACENET Data Catalog. 2018. Available online: https://www.epo.org/searching-for-patents/business/patstat.html#tab-4 (accessed on 14 May 2020).
5. Markellos, K.; Markellou, P.; Mayritsakis, G.; Perdikuri, K.; Sirmakessis, S.; Tsakalidis, A. Knowledge discovery in patent databases. In Proceedings of the 11th international conference on Information and knowledge management, McLean, WV, USA, 4–9 November 2002; pp. 672–677.
6. Ampornphan, P.; Tongnam, S. Patent knowledge discovery using data analytics. In Proceedings of the ICIT: International Conference on Information Technology, Singapore, 27–29 December 2017; pp. 42–46.
7. Larose, D.; Larose, C. *Discovering Knowledge in Data*; John Wiley Sons, Inc.: Hoboken, NJ, USA, 2014.
8. Zhuang, L.; Li, L.; Li, T. Patent mining: A survey. *ACM SIGKDD Explor. Newslett.* **2014**, *16*, 1–19. [CrossRef]
9. Ma, J.; Porter, A. Analyzing patent topical information to identifying technology pathways and potential opportunities. *Scientometrics* **2015**, *102*, 811–827. [CrossRef]
10. Jun, S.; Park, S.; Jang, D. Patent management for technology forecasting: A case study of the bio-industry. *JIPR* **2012**, *17*, 539–546.
11. Park, S.; Lee, S.-J.; Jun, S. A network analysis model for selecting sustainable technology. *Sustainablility* **2015**, *7*, 13126–13141. [CrossRef]
12. Choi, J.; Jang, D.; Jun, S.; Park, S. A predictive model of technology transfer using patent analysis. *Sustainablility* **2015**, *7*, 16175–16195. [CrossRef]
13. Choi, D.; Song, B. Exploring technological trends in logistics: Topic modeling-based patent analysis. *Sustainablility* **2018**, *10*, 2810. [CrossRef]
14. Liu, W.; Tao, Y.; Yang, Z.; Bi, K. Exploring and visualizing the patent collaboration network. *Sustainability* **2019**, *11*, 465. [CrossRef]
15. Witten, H.; Frank, E.; Hall, A. *Data Mining*; Morgan Kaufmann: Burlington, MA, USA, 2011.
16. Melvin, C.; Lee, W. Evaluation and improvement of procurement process with data analytics. *IJACSA* **2015**, *6*, 70–80.
17. Talib, A.; Hanif, M.; Ayesha, S.; Fatima, F. Text mining: Techniques, applications and issues. *Int. Adv. Comput. Sci. Appl.* **2016**, *7*, 414–418. [CrossRef]
18. Agrawal, R.; Srikant, R. Fast algorithms for mining association rules. In Proceedings of the 20th International Conference on Very Large Databases, San Francisco, CA, USA, 12–15 September 1994; pp. 478–499.
19. Yang, D.; Kang, J.; Park, Y.B.; Park, Y.J.; Oh, H.; Kim, S. Association rule mining and network analysis in oriental medicine. *PLoS ONE* **2013**, *8*, e59241. [CrossRef] [PubMed]
20. Farooq, A.; Uzair, M.; Joyia, G.; Akram, U. Detection of influential nodes using social networks analysis based on network metrics. In Proceedings of the IEEE International Conference on Computing, Mathematics and Engineering Technologies, Sukkur, Pakistan, 3–4 March 2018.
21. Lee, S.; Cha, Y.; Han, S.; Hyun, C. Application of association rule mining and social network analysis for understanding causality of construction defects. *Sustainablility* **2019**, *11*, 618. [CrossRef]
22. Singh, S.; Thapar, V.; Bagga, S. Exploring the hidden pattern of cyberbullying on social media. In Proceedings of the International Conference on Computational Intelligence and Data Science, Gurgaon, India, 6–7 September 2019; pp. 1636–1647.