International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Study on the Development of a Core Patent Classification Model Using Improved Patent Performance Indicators

Youngho Kim[1], Sangsung Park[2], Junseok Lee[3], Jiho Kang*[4]

Submitted: 02/12/2021 Accepted : 10/01/2022

Abstract: A patent contains various information about a developed technology and is a form of Big data that receives millions of applications worldwide each year. Recently, there has been an increase in research that analyzes such patent Big data for use in R&D strategy establishment. Among these studies, a core patent classification is recognized as important because it can be used for a variety of management information. In the past, the core patent classification was performed qualitatively by some experts, but it was expensive and time consuming. To complement qualitative methods, quantitative methods using statistics and machine learning are being studied. Existing proposed methods utilize the quantitative indicators specified in the patent. However, quantitative indicators have different values for each elementary technology. If this characteristic is not reflected, an incorrect analysis result is produced. In addition, various values such as rights, technology scalability sustainable development, etc., must be considered in order to effectively classify core patents. In this paper, we propose an effective core patent classification model using improved patent performance indicators. The proposed model applies text mining and clustering to patent Big data to identify elementary technology and calculate improved patent performance indicators that reflect various values. Furthermore, a core patent classification model is constructed by learning various classification algorithms. In order to examine the practical applicability of the proposed model, experiments are conducted with patents registered in the USPTO. As a result of the experiment, the accuracy of three models trained with patent-improved performance indicators was high. Among them, k-nearest neighbors demonstrated the highest performance.

Keywords: *Patent big data, Core patent classification, Patent performance indicators, Machine learning*

1. Introduction

Recently, Big data is produced through smartification, in which most industries are integrated with intelligent information technology. Countries or companies can analyze this Big data to create various values and improve competitiveness [1].

A patents include information on technologies developed in various forms, such as drawings and text. In addition, millions of applications are filed around the world every year, and the number of applications is steadily increasing [2]. For this reason, recently, there have been more studies that recognize patents as Big data and utilize them for Research and Development (R&D) strategy establishment through analysis [3]. Research using patent Big data include technology trend identification, R&D pattern analysis, and core technology classification. Of these, core technology or core patent classification is important because it can be used in a variety of management information applications, such as identifying competitors, preventing conflicts, and establishing portfolio strategies. Core patents are major patents that affect many technologies in a particular field. For this reason, the core patent is a high value patent that can provide many benefits to the owner. The core patent can be derived for a number of technologies constituting a specific technology field, that is, Elementary Technology (ET). ETs represent a more detailed technical field.

Existing core patent classification methods are mainly performed qualitatively by experts. Qualitative methods are time consuming and costly, and biased results can be obtained depending on the analyst. In order to address the above issue, recently, methods for deriving a core patent by applying statistics and machine learning to patent Big data have been proposed [4-14]. The proposed methods mainly use Quantitative Indicators (QI) specified in the patent for analysis, such as citation and number of families. The patents with high QI values are of high quality and are likely to be core patents. However, since QI has different values for each ET, it is necessary to analyze by reflecting these characteristics. Currently, ET identification methods such as IPC codes are granted according to the patent judge's qualitative judgment, making it difficult to secure homogeneity. In addition, various values of patents such as rights, technology scalability, sustainable development, etc., must be reflected for effective core patent classification.

In order to reflect various values of patents, the Korea Institute of Patent Information (KIPI) developed a set of composite indicators using QI [15]. However, the developed composite indicators are calculated from a wide range of technical aspects and cannot be applied to individual patent units. To compensate for this issue, the Korea Intellectual Property Strategy Institute (KIPSI) proposed a

[1] *Machine Learning Big Data Institute, Korea University, Seoul – 02841, KOREA*
ORCID ID : 0000-0001-8302-9818
[2] *Department of Big Data Statistics, Cheongju University, Chungbuk – 28503, KOREA*
ORCID ID : 0000-0001-6804-2707
[3] *MICUBE Solution, Seoul – 06719, KOREA*
ORCID ID : 0000-0003-0491-9690
[4] *Machine Learning Big Data Institute, Korea University, Seoul – 02841, KOREA*
ORCID ID : 0000-0001-8148-2555
** Corresponding Author Email: kangmae@korea.ac.kr*

patent performance indicator (PI) applicable to individual patents [16]. Nevertheless, the proposed PI is calculated differently depending on the ET and the time of analysis. Furthermore, in the case of convergence technology, it is difficult to identify the technical field, so only domain experts can calculate the PI value.

In order to address the above issue, this paper efficiently identifies ET through text mining and machine learning techniques and grasps characteristics through visualization. The ET identification method proposed in this paper can be applied to various technical fields, and it is possible to draw objective results. In addition, it develops a method of calculating improved patent performance indicators (improved PI) and a quantitative core patent classification model that can reflect various values of patents. The proposed Improved PI can comprehensively reflect values such as Rights, Sustainable development, Technology scalability, Technology impact, and Market power of patents according to characteristics of each ET in the core patent classification. In order to examine the applicability of the proposed model, experiments are conducted using patents registered in the United States Patent and Trademark Office (USPTO). The collected patents are related to the Medical robotics technology, with a total of 1,306 patents.

2. Related Studies

2.1. Patent Big Data Analysis

Big data is a concept that started from the technical aspect of massive amounts of data with the recent convergence of intelligent information technology. Generally, Big data satisfies 3V, which means high volume, velocity, and variety [17].

Millions of patents are applied for annually in various countries, and new technologies are mostly published as patents. Additionally, these patents contain structured or unstructured data such as text and drawings to contain detailed information on the developed technology. As such, the patent satisfies all 3V of Big data. Therefore, recent research has been conducted to recognize patents as big data and apply various analysis methods [18-23].

Lee et al. [18] used patent Big data to analyze the technology convergence pattern. To this end, the International Patent Classification (IPC) code was extracted from the collected data and association rules and network analysis were applied. Seo et al. [19] proposed a platform that utilizes existing Big data tools to effectively analyze patent Big data. Segev et al. [20] performed a regression analysis by extracting keywords of patents for technology trend analysis. Khoury and Bekkerman [21] proposed a method for measuring semantic similarity in patent Big data for prior art search. Qu et al. [22] recognized patents as Big data in science technology and measured keyword rank through the TF-IDF technique. Pilkington et al. [23] extracted and analyzed QI from patent Big data to understand the trend of technology development. As in the above studies, various values can be created by applying Big data analysis techniques such as statistics and machine learning to patents. In this paper, we propose a core patent classification model that applies Text mining, Visualization, and Machine Learning techniques to patent big data.

2.2. Core Patent Classification Using Patent Big Data

The core patent has a great influence within a specific technology field and creates a lot of profit. Such core patents can be used as a variety of information, such as identifying competitors, preventing conflicts, and establishing portfolio strategies. In general, core patents have high quality, and various values can be created. For this reason, companies conduct R&D, technology transfer, etc. to own core patents. In order to efficiently preempt a core patent, it is important to identify high-quality patents faster than competitors. Therefore, various methods have been proposed to effectively classify core patents [4-14].

Kim [4] extracts IPC codes from collected patents to perform Social Network Analysis (SNA) and derives the core technology field. Furthermore, patents with a large number of citations and families in the derived technology field were selected as core patents. The method proposed by Kim [4] uses only IPC codes that appear in collected patents, making it difficult to identify convergence or new technologies. In addition, various values of patents cannot be considered by using only the number of citations and families, respectively. Lee et al. [5] applied Principal Component Analysis (PCA) and Logistic Regression to the text of the patent and identified key technologies with significant keywords. However, the method proposed by Lee et al. [5] is difficult to apply to individual patent units. In addition, rights and technology impact cannot be measured using only text information. Wu et al. [6] proposed a method of extracting various QI and measuring the quality of patents through a Self-Organizing Map (SOM). However, QI has different values for each ET, but the method proposed by Wu et al. [6] does not take these characteristics into account. Woo et al. [7] proposed a core patent selection method through the Bayesian structural equation model. They qualitatively derived the criteria for selecting core patents through Delphi techniques. Qualitative methods are time consuming and costly, and it is possible that biased results may be obtained depending on the analyst. Wang et al. [8] proposed a method to perform SNA and select core patents using citation information of patents. Wu et al. [9] applied K-Means clustering to the Pearson correlation coefficient matrix, constructed using citation information to select core patents. The methods proposed by Wang et al. [8] and Wu et al. [9] used only citation information. Since the citation of patents has characteristics that are higher in older patents, various QI considerations are necessary. In this paper, improved PI is calculated using QI and PCA, which can measure various values, and a core patent classification model is constructed.

Trappey et al. [10] extracted important indicators using PCA and Kaiser-Meyer-Olkin (KMO) for various QI and trained them with a neural network to select core patents. In addition, Trappey et al. [10] used IPC codes to reflect differences in QI values for each ET. Cho and Shih [11] used IPC codes to identify core technology at the national level. However, in Trappey et al. [10] and Cho and Shih [11], the IPC codes that were used were given according to the previously disclosed technology fields, making it difficult to identify convergence and new technology fields. Chang [12], Lee, and Su [13] and Chang [14] used Cooperative Patent Classification (CPC) codes to identify technological trends and select core technologies. CPC codes can identify technological development trends from a macroscopic perspective, but it is difficult to identify ET. In addition, studies using CPC codes are technical unit analysis methods and are difficult to apply to individual patents. In this paper, ET is quantitatively identified by using text information of patents, so it can be applied to various technical fields and be used for individual patent units through Improved PI.

KIPI [15] proposed composite indicators using patented QI to identify technological innovation activities. The proposed composite indicators are calculated through the average number of cited patents by company or technology. Therefore, it can be analyzed by a specific company or technology unit, but it is difficult to apply to individual patents. To complement this, KIPSI [16] proposed a PI applicable to individual patents as follows:

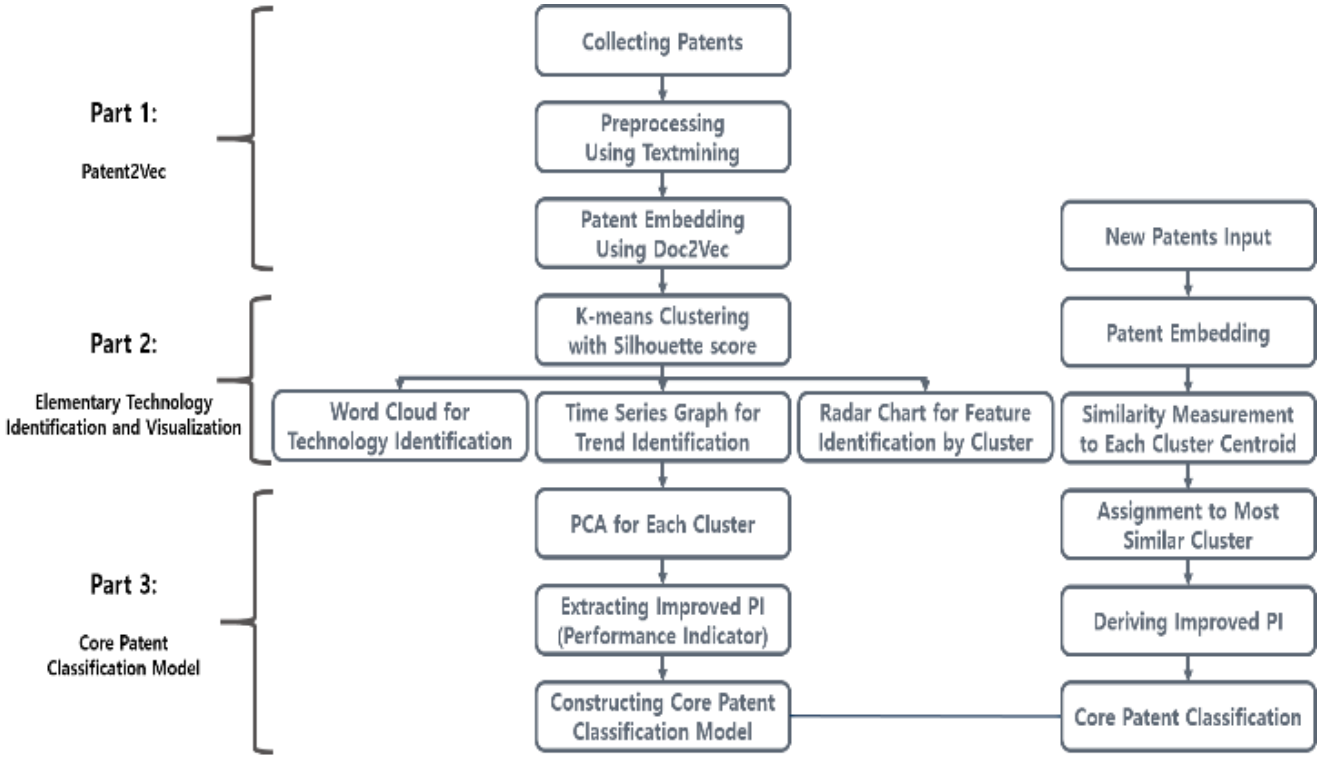

Fig. 1. Proposed methodology

$$PI(i) = \frac{p_i}{\frac{p_1+\cdots+p_n}{n}} \quad (1)$$

In the Equation (1), $PI(i)$ is the PI value of the *i*-th patent, while n is the number of patents with the same registration year and technical classification among the collected patents. In addition, p_n is the QI of the *n*-th patent belonging to the same registration year and technical classification, and p_i represents the QI value of the patent to be analyzed. $PI(i)$ is applicable to most patents QI, such as the number of citations, number of claims, and number of families. However, PI is calculated differently depending on the analysis time and technical classification. Furthermore, technical classification mainly uses IPC codes, but it is difficult to identify effective ET. This is because the IPC code is multi-classified, the technical scope is wide, and subjective judgment of the patent examiner is applied. To solve this issue, this paper effectively identifies ET by using the text mining technique. Moreover, we propose an improved PI that can be applied to various technical fields by extracted key variables through applying PCA for each ET.

2.3. PCA for deriving improved PI

PCA creates new variables that are orthogonal to each other by linearly transforming the original variables while preserving the variance of the data as much as possible [24]. PCA is the most widely used dimensionality reduction technique by transforming data in a high-dimensional space into an uncorrelated low dimension. PCA creates the same number of principal components (PCs) as the dimensions of the original variables, and PC is derived using data values and factor loadings as below:

$$\begin{aligned} PC_1 &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ PC_2 &= a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ &\vdots \\ PC_m &= a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \end{aligned} \quad (2)$$

In the Equation (2), PC_m represents the *m*-th principal component, x_n is the original data value, and a_{mn} represents factor loadings. As in equation (2), factor loadings indicate the importance of each variable [25]. PCA extracts variables to better represent data variance. For this reason, this paper performs PCA for each ET and derives variables with high explanatory power, that is, improved PI. Moreover, the core patent classification model is constructed using improved PI as an input variable. The Improved PI proposed in this paper is derived by performing PCA for each ET. The proposed method can reflect characteristics with different indicator values for each ET. In addition, Right, Sustainable development, Technology scalability, Technology impact, and Market power can be comprehensively considered in individual patents by deriving new variables using various Quantitative indicators.

3. Proposed methodology

In this paper, we propose a core patent classification model using ET identification and improved PI. The proposed model consists of three parts. First, in Part 1, patent Big data is converted into an appropriate form for analysis for ET identification. “Abstract,” which is the text information of the patent, is extracted and preprocessed through text mining. In addition, patent Big data is vectorized (Patent2Vec) using Doc2Vec that can embed documents in a specific space while preserving context

information. In part 2, ET is identified by clustering. For this, K-Means clustering is used and the optimal number of clusters is selected through the Silhouette score. When clustering is used to identify ET, quantitative and objective results can be derived based on the technical similarity of patents. Moreover, in order to understand the characteristics of ETs, it is visualized through Word Cloud, Time Series Graph, and Radar Chart. In Part 3, PCA is used to extract improved PI for each ET. It derives Improved PI by performing PCA on Quantitative indicators of patents for each identified ET. Finally, a core patent classification model is constructed using improved PI as input variables. In addition, the core patent label for constructing a classification model is given as whether the technology is transferred. The identification of ET when new data input is as follows. First, Patent2Vec is performed on the abstract of patent. Next, measure similarity to the existing identified ETs. Assign patent to the most similar ET and derive Improved PI. Finally, it is applied to the constructed classification model. Figure 1 is a schematic diagram of the proposed methodology in this paper.

3.1. Data Description

In this paper, we propose an effective core patent classification model using patent Big data. In order to examine the applicability of the proposed model, experiments are conducted by collecting actual patents. A total of 1,306 patents related to medical robotics technology were collected and registered in the USPTO. The data collection process is as follows. First, derive keywords related to the field of medical robotics to create a draft search formula. Next, the search process is performed several times, synonyms are added, and a final search formula is prepared. Data is collected using a search formula and duplicate data is removed to obtain final analysis data. The reason USPTO was selected as the data collection DB in this paper is that the most patents in the medical robotics field are applied in the United States. In addition, USPTO's patent data is open to the public for easy collection. The data collection period is from 2002 to 2018. The core patent label for constructing a classification model is given as to whether the patent has been transferred. This is because, in general, technology transfer involves selling patents with excellent quality to consumers [10]. Table 1 summarizes the characteristics of the experimental data.

Table 1. Information on experimental data

Technological field	Database	Period	Status	Number of patent (Number of transferred)
Medical robotics	USPTO	2002~ 2018	Registered	1,306 (365)

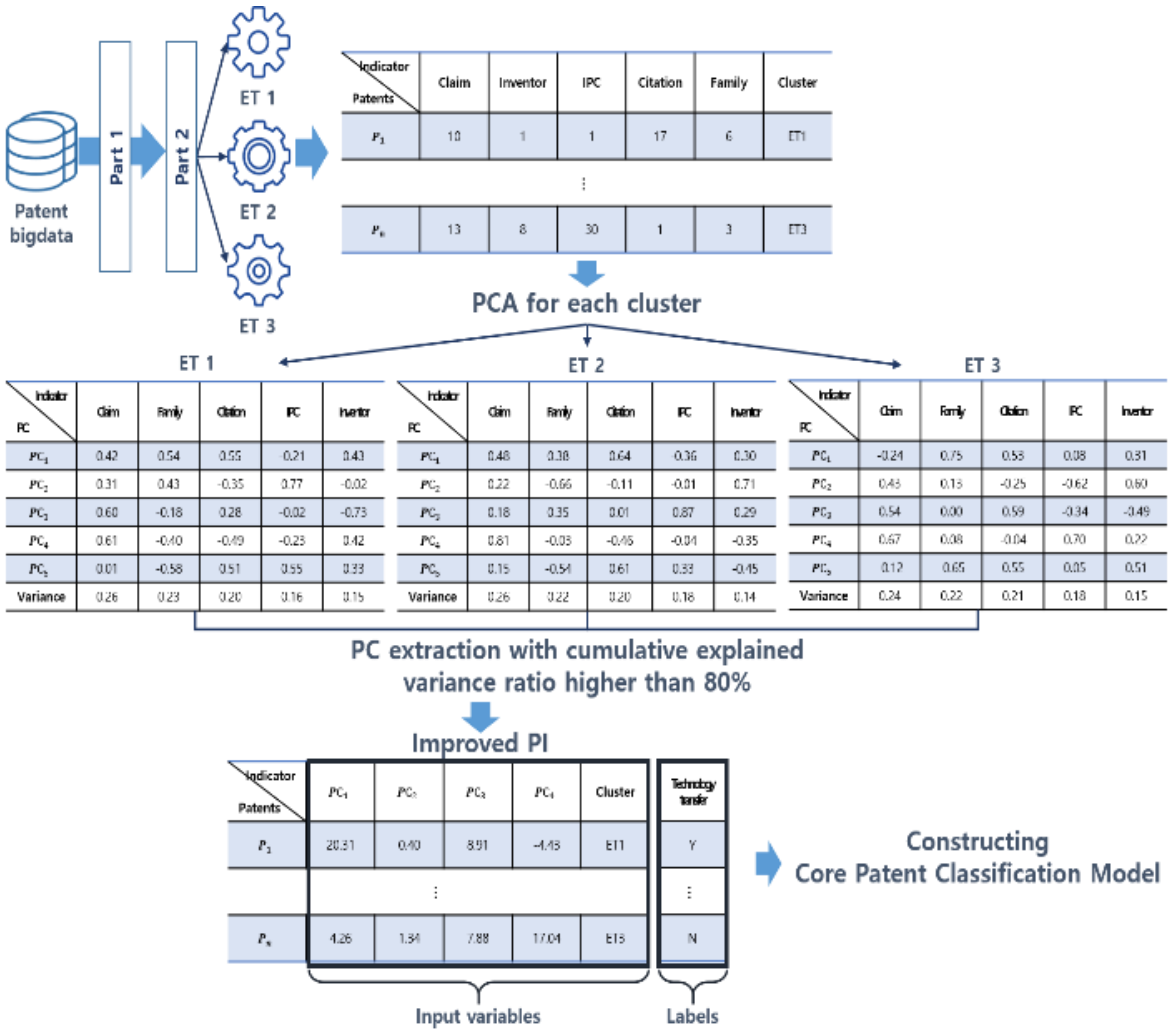

Fig. 2. Improved PI calculation process

3.2. Patent2Vec

The patent is data that includes both structured and unstructured forms, statistics and machine learning techniques cannot be applied directly. Therefore, in order to quantitatively analyze the patent, it must be converted into an appropriate form.

In this paper, abstracts containing technical information of patents are extracted and vectorized. To perform vectorization, lemmatization, which is effective in extracting the basic form of words, is used among text mining techniques. The lemmatization removes the ending of a word and can match the form of a tense. For this reason, lemmatization is often used as a preprocessing technique in text mining. Table 2 is an example of lemmatization of the abstract part of the patent.

Table 2. Example of lemmatization

Methods	Sentence
Raw data	A robotic arm including a parallel spherical five-bar linkage with a remote center of spherical rotation
Lemmatization	Robot include parallel linkage remot center spheric rotat

The abstract part of the patent contains some meaningless words in the analysis such as 'the' and 'a.' These are called stop words and are eliminated because they reduce the efficiency of the analysis. The preprocessed text information is learned with Doc2Vec, a document-embedding algorithm based on neural networks. Doc2Vec preserves context information and embeds it in a specific space, thus making it more effective than the existing Bag of Words (BOW) method [26], [27], [28]. In the proposed model, the patent is embedding in a 200-dimensional space using Distributed Memory Model of Paragraph Vectors (PV-DM), which is known to have excellent performance among Doc2Vec.

3.3. Elementary Technology Identification and Visualization

IPC, CPC codes, etc., which indicate the fields of related technologies, are assigned to patent documents. However, most of the patents are new technologies that have not been released to the public, and recently, it has become difficult to assign appropriate codes due to convergence between technologies.

In this paper, clustering is applied to Patent2Vec performed in Part 1 to effectively identify ET. Clustering is performed so that similar documents are grouped in the same cluster according to data characteristics [29]. Therefore, ET can be effectively identified in patent Big data. The clustering method uses K-Means clustering, which is typically used. In addition, as can be seen in Equation (3), the optimal number of clusters—that is, the number of ET—is quantitatively selected through the silhouette score.

$$s_i = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}, Silhouette\ score = \sum_{i=1}^{n} s_i \quad (3)$$

In Equation (3), $a(i)$ is the average of the distances between elements in the cluster to which the individual i-th element belongs; $b(i)$ is the average distance between i-th element and elements belonging to another cluster. In general, the number of clusters is selected as the largest silhouette score, but if the number of clusters is too large, it is difficult to understand the results of clustering.

Following the clustering, visualization techniques are used to define technical characteristics and identify each ET. First, the word cloud can define ET by expressing words of patents included in each cluster. Next, to grasp the development trend by ET, a time series graph is created using the application date of the patent. Finally, a radar chart is made for each ET with the extracted QI, as can be seen in Table 3, and the characteristics are identified.

Table 3. Patent QI for creating radar chart

Quantitative indicators (QI)	Description	Measurable value information
Claim	Number of claims	Rights
Inventor	Number of inventors	Sustainable development
IPC	Number of IPC codes	Technology scalability
Citation	Number of forward citation	Technology impact
Family	Number of family nations	Market power

In Table 3, “Claim” represents the scope of technology protection of the patent, so it is possible to measure the value of rights. “Inventor” represents the number of researchers that invested in developing the technology. Therefore, since the accuracy and fidelity of the invention can be judged, the degree of sustainable development can be measured. “IPC” is a code that is related to the relevant patent among existing technical fields. Accordingly, the more IPC codes assigned, the more scalable it is with various technologies. “Citation” is the degree to which the patent is cited by other patents, so it is possible to measure the technological impact. Finally, “Family” represents the degree to which the same technology has been applied in various countries. Therefore, it is possible to measure market power.

3.4. Core Patent Classification Model

In order to effectively classify core patents in patent Big data, it is necessary to reflect various values. The model proposed in this paper uses the improved PI derived by applying PCA to the QI extracted in Part 2. Therefore, it is possible to reflect various values. In addition, since ET is identified based on data, it can be applied to various technical fields and is possible to classify core patents accordingly.

In Part 3, PCA is performed to derive the improved PI and the cumulative explained variance ratio is checked. To this end, PCA is performed for each ET identified in Part 2 to extract principal components. PCA can extract new variables for each ET. Improved PI extracted through this process can measure various values of patents by ET. In addition, a core patent classification model can be built using the improved PI as an input variable. Figure 2 shows the improved PI calculation process.

Improved PI is trained with various classification algorithms and performance is compared. Logistic Regression (LR), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and AdaBoost (AB) are used as classification algorithms for performance comparison. In addition, a grid search method is used to optimize parameters for each algorithm and a model showing generalized classification performance is constructed through 10-fold cross validation (CV). Table 4 summarizes the optimized parameters for each classification algorithm.

Table 4. Optimized parameters by algorithms

Algorithms	Parameters
LR	L2 penalty
KNN	K=4, Distance=Minkowski
SVM	Kernel=RBF, C=1, Gamma=10
DT	Criterion=Gini, Max_leaf_nodes=59
RF	Criterion=Gini, Number of trees=200, Number of features=2
AB	Criterion=Gini, Number of trees=50, Number of features=2

Accuracy, Precision, Recall and F1-score are used in this paper to measure core patent classification performance. Accuracy represents the degree of agreement between the actual label and the

predicted label. Since the core patent label is unbalanced, Precision, Recall, and F1-score are calculated, as can be seen in Table 5, and considered together.

Table 5. Equation of performance measure

Performance measure	Equation
Accuracy	$\frac{TP+TN}{P+N}$
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
F1-score	$2*\frac{Precision*Recall}{Precision*Recall}$

In Table 5, True Positive (TP) is the prediction of the actual number of positive labels. In addition, True Negative (TN) is prediction of the actual number of negative labels. False Positive (FP) and False Negative (FN) are mispredictions of actual labels.

4. Experiment Results

In order to verify the practical applicability of the core patent classification model proposed in this paper, experiments are conducted on medical robotics technology–related patents. First, Patent2Vec, which vectorizes the collected patents, is performed in Part1. Lemmatization was used to convert the patent abstract to a basic form. Additionally, as can be seen in Table 6, words frequently appearing in the medical robotics patent set are identified as stop words and removed.

Table 6. List of additional stop words in medical robotics patent set

Medical robotics stop words
'invent', 'method', 'system', 'includ', 'model', 'data', 'provid', 'apparatus', 'techniqu'

The preprocessed patent was embedding in a 200-dimensional space through Doc2Vec. K-Means clustering was performed to identify ET in the embedded patent. When performing clustering, the number of K was derived as the largest silhouette score as shown in Figure 3.

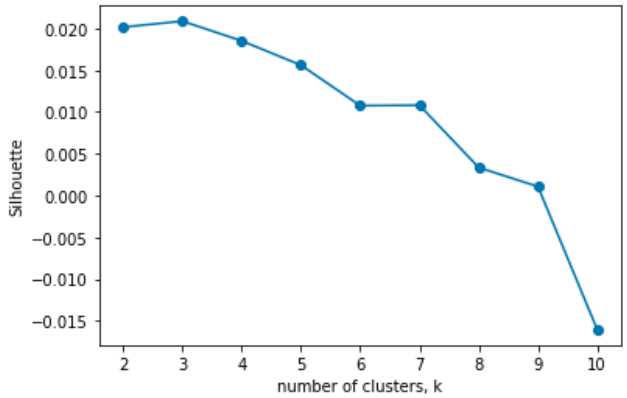

Fig. 3. Result of silhouette score

Figure 4 shows PCA performed on the Doc2Vec values of all patents to visualize the clustering results. In Figure 4, dots represent individual patents and the colour of the dots represents clusters.

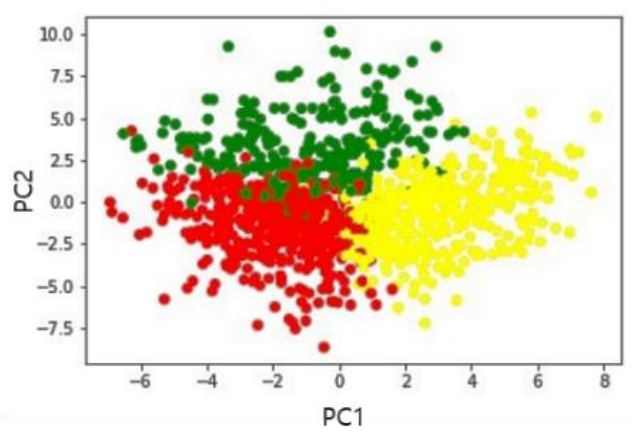

Fig. 4. Result of clustering

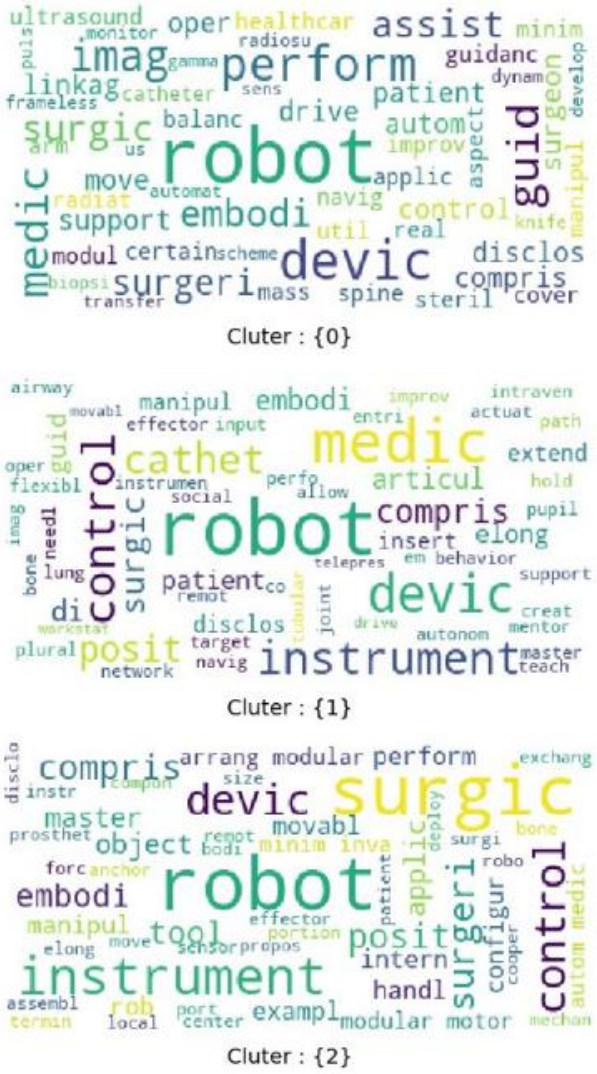

Fig. 5. Result of word cloud by cluster

Word cloud is used to check the technology of each cluster. Word cloud visualizes the words contained in the cluster, where words that appear more frequently are larger in size. Figure 5 shows the results of word cloud analysis by cluster.

In Figure 5, Cluster 0 was derived from the words 'robot,' 'perform,' 'devic,' 'guid' and 'assist.' Therefore, Cluster 0 is a surgery assistant robot technology that is a surgical tool that helps in surgical activities [30]. In Cluster 1, words such as 'robot,' 'medic,' 'control,' and 'posit' frequently appear. For this reason, Cluster 1 is a non-surgical robot technology that performs various

medical activities except surgery [31]. We know that Cluster 2 is a surgical robot technology that directly performs surgery because words such as 'robot,' 'surgic,' 'embodi,' and 'manipul' frequently appear [30], [31]. Table 7 summarizes technology definitions and frequently occurring words for each cluster.

Table 7. ET identification using word cloud

Cluster number	High frequency words	Technology definition
Cluster 0	'robot', 'assist', 'guid', 'perform', 'devic'	Surgery assistant robot
Cluster 1	'robot', 'medic', 'control', 'posit'	Non-surgical robot
Cluster 2	'robot', 'surgic', 'embodi', 'manipul'	Surgical robot

Next, to analyze the identified trends by ET, we created a time series graph, as can be seen in Figure 6. The time series graph is plotted by cluster using the application dates of patents. The x-axis represents the application year, and the y-axis is the number of patent applications by cluster.

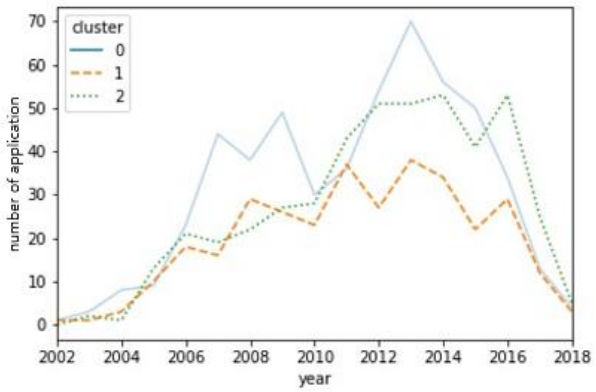

Fig. 6. Time series graph by cluster

In Figure 6, all clusters have a year-to-year increase or decrease, with the former being more common. Since 2010, patent applications of all clusters have increased rapidly. Cluster 0, which stands for surgery assistant robot technology, has recently increased rapidly. Accordingly, it appears that the application of the surgical robot technology utilized together increases. To understand the characteristics of each cluster, QI is extracted and a radar chart is created, as can be seen in Figure 7.

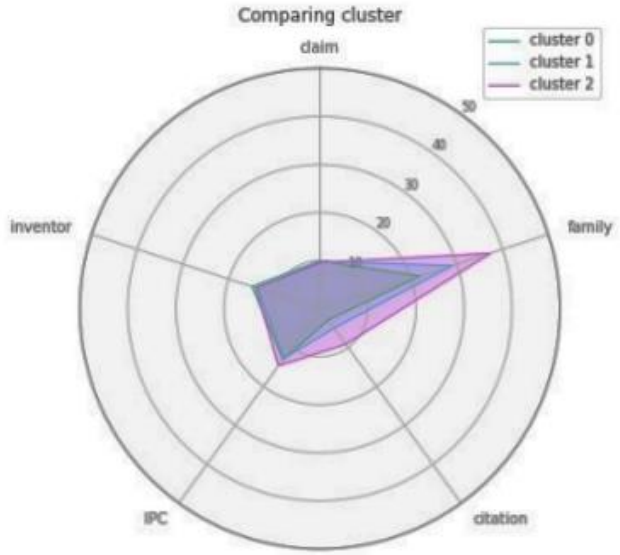

Fig. 7. Radar chart by cluster

In Figure 7, most of the medical robotics technologies are applied in various countries to secure market power. Cluster 2, which stands for surgical robot, has a higher citation than the other ETs, so many patents with high technology impact are included.

PCA uses ET and QI derived from the analysis results of Part 2. PCA was performed for each ET and the resulting variances by PC can be seen in Table 8.

Table 8. Variance ratio by cluster

Cluster number	Variance ratio				
	PC1	PC2	PC3	PC4	PC5
Cluster 0	0.26	0.23	0.20	0.16	0.15
Cluster 1	0.26	0.22	0.20	0.18	0.14
Cluster 2	0.24	0.22	0.21	0.18	0.15

As shown in Table 8, cumulative explained variance ratio of 80% or more among PC derived by applying PCA to QI is extracted. If the PCs are extracted and used when the cumulative explained variance ratio is 80%, most of the information in the original data can be preserved. Therefore, in the experiment, up to PC4 is derived as Improved PI. The improved PI is learned in an optimized classification algorithm to measure performance, and finally a core patent classification model is constructed. Figure 8 shows the performance of each classification algorithm.

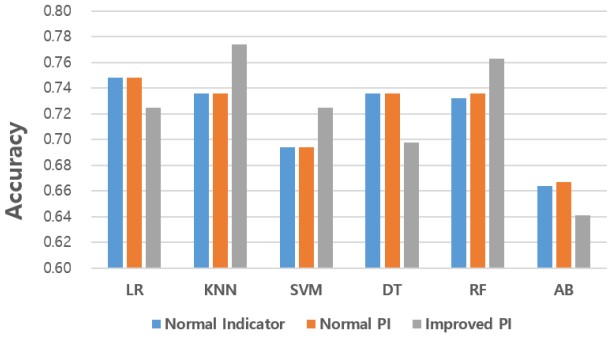

Fig. 8. Accuracy by input variables and models

In Figure 8, "Normal Indicator" is the QI extracted from data; "Normal PI" is the PI proposed by KIPSI; and "Improved PI" is a method proposed in this paper. It was found that the accuracy of the proposed method was high in KNN, SVM and RF. In particular, the proposed method has the highest performance in KNN. Figure 9 shows Precision.

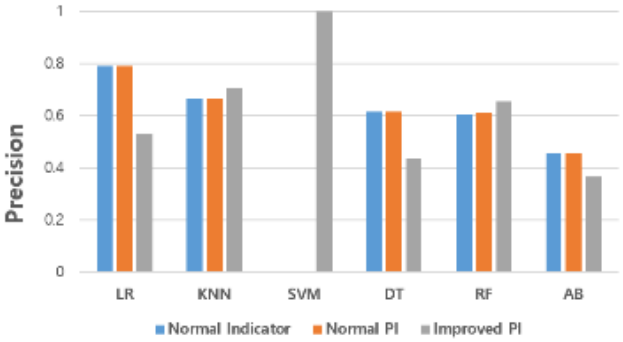

Fig. 9. Precision by input variables and models

In Figure 9, precision is mostly same except for SVM. Figure 10 shows Recall.

Recall is mostly derived with high values of "Normal Indicator" and "Normal PI". Figure 11 shows F1-Score. In Precision, Recall, and F1-score, the value of indicator was mostly derived as 0. This is because SVM performed prediction with only one label. As for the F1-Score, the KNN of the proposed method is derived the

highest as the accuracy result.

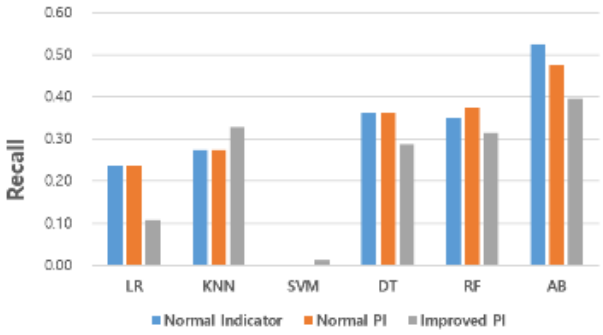

Fig. 10. Recall by input variables and models

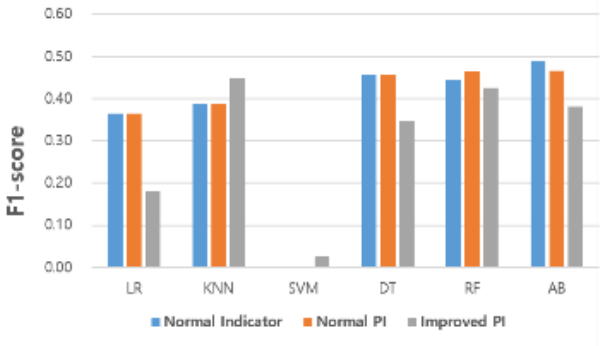

Fig. 11. F1-Score by input variables and models

5. Conclusion and Future Work

The existing research on the classification of core patents does not identify ET or is difficult to apply to individual patent units. Also, it does not reflect various values using only a specific QI. To solve this issue, this paper proposed an effective core patent classification model using patent Big data. The proposed model can analyze various technical fields by quantitatively identifying ET through text mining and clustering and identifying characteristics through visualization. In addition, it derives the improved PI and learns the classification algorithm to reflect the features and various values of each ET.

In order to examine the practical application of the proposed model, experiments were conducted by collecting medical robotics–related cases among USPTO registered patents. As a result of the experiment, three ETs could be identified and technical definitions and characteristics could be identified through visualization. The three ETs constitute a medical robot technology and include Surgery assistant robot, Non-Surgical robot, and Surgical robot. Furthermore, as a result of learning and testing various classification algorithms by deriving the improved PI, the proposed model demonstrated high performance.

Technology has a life cycle of introduction, diffusion, and decline. For this reason, the identification result of the core patent for each technology may vary depending on the time of analysis. The current technology life cycle analysis is visually expressed to identify only trends. This method of visualizing the technology life cycle is effective for identifying trends from a macroscopic perspective. However, it is difficult to identify the core patent by life cycle. In future works, it will be necessary to study how to reflect the features of technology life in the identification of core patents. To this end, large-scale data collection and visualization must be performed to sufficiently grasp the technology life cycle. In addition, a method for extracting features of patents applied for each life cycle should be studied.

Acknowledgements

This research was supported by the MOTIE (Ministry of Trade, Industry, and Energy) in Korea, under Human Resource Development Program for Industrial Innovation (Global) (P0017311) supervised by the Korea Institute for Advancement of Technology (KIAT).

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] K. Witkowski, “Internet of things, big data, industry 4.0 – innovative solutions in logistics and supply chains management,” *Procedia Eng.*, vol. 182, pp. 763–769, 2017.

[2] World Intellectual Property Organization, “World intellectual property indicators 2019,” *WIPO.*, vol. 941E, no. 19, pp. 1–224, 2019.

[3] L. Zhang, L. Li and T. Li, “Patent mining: a survey,” *SIGKDD Explor. Newsl.*, vol. 6, no. 2, pp. 1–19, 2015.

[4] H. Kim, “A novel methodology for extracting core technology and patent by ip mining,” M.S. thesis, Dept. Ind. Manag. Eng., Korea Univ., SU, Korea, 2016.

[5] J. Lee, J. LEE, G. Kim, S. Park and D. Jang, “Establishment of strategy for management of technology using data mining technique,” *J. Korean Inst. Intell. Syst.*, vol. 25, no. 2, pp. 126–132, 2015.

[6] J. Wu, P. Chang, C. Tsao and C. Fan, “A patent quality analysis and classification system using self-organizing maps with support vector machie,” *Appl. Soft Comput.*, vol. 41, pp. 305–316, 2016.

[7] H. Woo, J. Kwak and C. Lim, “A study on patent evaluation model based on bayesian approach of the structural equation model,” *Korean J. Appl. Statist.*, vol. 30, no. 6, pp. 901–916, 2017.

[8] C. Wang, C. Chiang and S. Lin, “Network structure of innovation: can brokerage or closure predict patent quality?,” *Scientometrics*, vol. 84, pp. 735–748, 2010.

[9] H. Wu, H. Chen and K. Lee, “Unveiling the core technology structure for companies through patent information,” *Technol. Forecasting Soc. Change*, vol. 77, no. 7, pp. 1167–1178, 2010.

[10] A. J. C. Trappey, C. V. Trappey, C. Wu and C. Lin, “A patent analysis for innovative technology and product development,” *Adv. Eng. Inform.*, vol. 26, no. 1, pp. 26–34, 2012.

[11] T. Cho and H. Shih, “Patent citation network analysis of core and emerging technologies in Taiwan: 1997-2008,” Scientometrics, vol. 89, pp. 795, 2011.

[12] S. Chang, “Key technologies and development trends of 5g optical networks,” *Appl. Sci.*, vol. 9, no. 22, 2019.

[13] M. Lee and W. Su, “Search for the developing trends by patent analysis: a case study of lithum-ion battery electrolytes,” *Appl. Sci.*, vol. 10, no. 3, 2020.

[14] S. Chang, “Patent analysis of the critical technology network of semiconductor optical amplifiers,” *Appl. Sci.*, vol. 10, no. 4, 2020.

[15] R&D Intellectual Property Information System, SU, Korea. Patent analysis methodology for making of technology roadmap. [Online]. Available: http://www.ripis.or.kr/U_Pds.do?method=m011&ntcbd_mng_seq=12&wrt_seq=336

[16] R&D Intellectual Property Information System, SU, Korea. Guidelines for utilizing patent performance indicators. [Online]. Available: http://rndip.or.kr/U_Pds.do?method=m011&ntcbd_mng_seq=12&wrt_seq=345

[17] R. Lu, H. Zhu, X. Liu, J. K. Liu and J. Shao, “Toward efficient and privacy-preserving computing in big data era,” *IEEE Netw.*, vol. 28,

no. 4, pp. 46–50, 2014.

[18] W. S. Lee, E. J. Han and S. Y. Sohn, "Predicting the pattern of technology convergence using big-data technology on large-scale triadic patents," *Technol. Forecasting Soc. Change*, vol. 100, pp. 317–329, 2015.

[19] W. Seo, N. Kim and S. Choi, "Big data framework for analysing patents to support strategic r&d planning," presented at DASC. *PiCom. DataCom. CyberSciTech.*, Auckland, New Zealand, Aug. 8-12, 2016.

[20] A. Segev, C. Jung and S. Jung, "Analysis of technology trends based on big data," presented at *2013 BIGDATA-CONGRESS*, Santa Clara, USA, Jun. 27-Jul. 2, 2013.

[21] A. H. Khoury and R. Bekkerman, "Automatic discovery of prior art: big data to the rescue of the patent system," *J. Marshall Rev. Intell. Prop. L.*, vol. 16, pp. 45–65, 2016.

[22] P. Qu, J. Zhang, C. Yao and W. Zeng, "Identifying long tail term from large-scale candidate pairs for big data-oriented patent analysis," *Concurr. Comp-Pract. E.*, vol. 28, pp. 4194–4208, 2016.

[23] A. Pilkington, R. Dyerson and O. Tissier, "The electric vehicle:: patent data as indicators of technological development," *World Pat. Inf.*, vol. 24, no. 1, pp. 5–12, 2002.

[24] J. Shlens, "A tutorial on principal component analysis," *arXiv*, preprint arXiv:1404.1100, 2014.

[25] Y. Chao and C. Wu, "Principal component-based weighted indices and a framework to evaluate indices: results from the medical expenditure panel survey 1996 to 2011," *PLoS One*, vol. 12, no. 9, e0183997, 2017.

[26] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient estimation of word representations in vector space," *arXiv*, preprint arXiv:1301.3781, 2013.

[27] X. Rong, "Word2vec parameter learning explained," *arXiv*, preprint arXiv:1411.2738, 2016.

[28] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents," *arXiv*, preprint arXiv:1405.4053, 2014.

[29] A. Sharma, "A survey on different text clustering techniques for patent analysis," IJERT., vol. 1, no. 9, pp. 1–4, 2012.

[30] J. E. Speich and J. Rosen, "Medical robotics," in *Encyclopedia of Biomaterials and Biomedical Engineering*, G. Wnek, G. Bowlin, Ed. Boca Raton, FL, USA: CRC Press, 2008, pp. 983–993.

[31] H. F. M. Van der Loos, D. J. Reinkensmeyer and E. Guglielmelli, "Rehabilitation and Health Care Robotics," in *Springer Handbook of Robotics*, B. Siciliano, O. Khatib, Ed. Cham, ZG, Switzerland: Springer, 2016, pp. 1685–1728.