International Journal of Engineering & Technology, 7 (3) (2018) 1501-1503

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i3.12487
Research paper

Scientific and industrial keyword analysis using structured covariance and clustering

Sunghae Jun [1] *, Seung-Joo Lee [1]

[1] Department of Big Data and Statistics, Cheongju University, Chungbuk 28503, Korea
**Corresponding author E-mail: statcs@gmail.com*

Abstract

Using scientific and industrial keyword analysis (SIKA), many industrial companies have built their research and development (R&D) strategies for improving technological competitiveness in market. The technological keywords extracted from journal papers and patent documents are good resources for SIKA. In this paper, we use patent keyword data as scientific and industrial keywords for SIKA. A patent contains various information of developed technology such as patent title, abstract, date, citation, etc. Because the exclusive rights of technologies applied and registered to patent system are protected by patent law for a certain period. We also consider statistical methods for the SIKA. First we perform technology clustering using K-means clustering of technological patent keywords. Next we carry out the principal component analysis (PCA) from the clustering results. Using the first and second principal components, we obtain PCA plots for techno-logical clusters. So we can understand the technological structure of given and target technology from the PCA plot results. Combing the technology clustering and PCA plots, we propose a method of SIKA to build valuable R&D strategy of company. To illustrate how the proposed method could be applied to real problem, we make experiments using many technological keywords for given technology field.

Keywords: *Scientific and Industrial Keywords; Structured Covariance Model; Principal Component Analysis; Technology Clustering.*

1. Introduction

It is important to know the industrial system for sustaining and improving competitiveness of company. One of the ways to manage industrial systems is to analyze technology that makes up the industrial systems. Technology analysis is to analyze technological resources related to target technology in industry fields. Many companies have been conducting scientific and industrial keyword analysis (SIKA) for technology management such as research and development (R&D) planning [1 - 3]. We have understood diverse technological structure from the SIKA results. In the most SIKA tasks, technological keywords extracted from patent document data have been used for technology analysis [4 - 5]. Because patent contains various and rich information about developed technologies such as title, abstract, inventors' name, citation, international patent classification, etc. [6]. So, in this paper, we analyze the technological keywords from patent document data as scientific and industrial keywords. We retrieve patent documents related to target technology area from patent databases, and extract technological keywords from the collected patent documents. Using the technological keywords, we perform the SIKA for understanding technological structure between sub technologies. Also we combine technology clustering and structured covariance analysis of patent keywords for the proposed SIKA. Lastly we carry out a case study to show how our study could be applied to practical domain using patent documents related to target technology. In this case study, we consider light emitting diode (LED) technology as a target technology, because this technology is one of important technologies for our life. Next section shows our proposed method for SIKA. We carry out the case study to verify our proposed method in section 3. In the last section, we conclude our research and provide the contribution of this paper.

2. Structured covariance clustering for SIKA

To carry out the SIKA, we should use statistical methods or machine learning algorithms. In the SIKA works, many researchers have tried to perform technological keywords analysis for their R&D planning. In this paper, we also propose an analytical method of technological keyword analysis for SIKA, and combine structured covariance analysis and clustering for the proposed method. We consider principal component analysis (PCA) and K-means clustering for structured covariance analysis and technology clustering respectively. To verify the performance of our research, we make experiments using the technological keywords related to the LED technology. We propose a method of technology keyword analysis for SIKA. This consists of PCA and K-means clustering from statistics and machine learning. The PCA is a statistical method of data compression using linear algebra [7]. The data set of SIKA consists of a matrix with n rows and p columns as follow.

$$(X_{i1}, X_{i2}, \dots, X_{ip}), i = 1,2, \dots, n \quad (1)$$

Where the row and column of the matrix represent patent document and technological keyword respectively. PCA provides a linear combination of principal components (PCs) using their covariance structure [8]. The jth PC is represented as follow.

$$PC_j = L_{j1}X_1 + L_{j2}X_2 + \cdots + L_{jp}X_p, j = 1,2, \dots, k \quad (2)$$

Where k is less than or equal to p. In the result of PCA, the first principal component (PC1) contains the most information about the entire keyword. Next, the second principal component (PC2) contains much information. In general, we use the first one, two, or three PCs to represent all technological keywords [8]. Using the PC1 and PC2, we understand the technological structure of given technological domain. The plot shows the various information hidden in the data of technological keywords. We apply this PCA plot to technology clustering for SIKA. In our study, we consider K-mean clustering to get the technology clustering results for SIKA. K-means clustering is a cluster analysis to group all object (documents) to similar clusters by distance measure between objects [9]. For using K-means clustering, we find the K value which is optimal number of clusters for give data set. In the previous researches related to technology clustering, the Silhouette measure were used for selecting the number of clusters [10 - 11]. In the Silhouette measure, we compute the standardized distance between each object and all other objects (A) and another distance between each object and all objects in the nearest cluster (B). Using the difference between A and B, the Silhouette measure provides the optimal number of clusters [12]. We determine the number of clusters with the largest Silhouette value. The proposed method for SIKA is carried out by the following steps.

Step 1: Data set preparation

(1. 1) Collecting technological documents related to target technology

(1, 2) Extracting keywords from collected document data

(1.3) Constructing matrix with document (row) and keyword (column)

Step 2: Technology clustering

(2.1) Finding optimal number of clusters (K) using Silhouette measure

(2.2) Performing K-means clustering using constructed matrix

(2.3) Assigning all documents to K clusters by K-means clustering result

Step 3: PCA plotting

(3.1) Performing PCA according to K clusters

(3.2) Showing PCA plots with first and second PCs

(3.3) Interpreting all PCA plots from technology perspective

We summarize the proposed method in the three steps. From the patent documents searching and preprocessing for PCA and clustering, we use the R data language and its packages [13,14]. Next we show how our method could be applied to practical problems related to SIKA.

3. Experimental results

In this section, we carried out various experiments to verify the performance of our proposed research. From the patent databases [15 - 16], we collected the patent documents related to the LED technology for our experiments. We searched all LED patents from 1973 to 2016, and the total number of collected patents was 4,206. For performing the proposed method, we extracted the patent keywords representing the LED technology from the patent documents as follows; ‘accord’, ‘adjust’, ‘arrang’, ‘board’, ‘bodi’, ‘chip’, ‘circuit’, ‘color’, ‘communic’, ‘connect’, ‘control’, ‘detect’, ‘devic’, ‘diod’, ‘direct’, ‘display’, ‘electr’, ‘electrod’, ‘element’, ‘emit’, ‘energi’, ‘fix’, ‘fluoresc’, ‘heat’, ‘illumin’, ‘inform’, ‘integr’, ‘intellig’, ‘lamp’, ‘layer’, ‘light’, ‘manufactur’, ‘materi’, ‘modul’, ‘oper’, ‘optic’, ‘organ’, ‘panel’, ‘plate’, ‘power’, ‘process’, ‘radiat’, ‘receiv’, ‘reflect’, ‘remot’, ‘screen’, ‘sensor’, ‘signal’, ‘structur’, ‘suppli’, ‘surfac’, ‘switch’, ‘temperatur’, ‘termin’, ‘time’, ‘voltag’, and ‘wireless’. To extract the keyword related to LED, we had been helped from the domain experts. So we constructed the data set (matrix) consisting of 4,206 documents (rows) and 57 keywords (columns). First, we performed the PCA on the structured data set. At this time, the number of principal components (PCs) was two and three[7]. In this paper, we decided the number of PCs to be two for visualization of PCA plots according to clusters. To select the optimal number of clusters, we computed the Silhouette widths, and we calculated the Silhouette values when the number of PCs was 2 and 3 according to the number of cluster. Since the Silhouette width is largest when the number of clusters is 3, we decided that the optimal cluster numbers for LED technology cluster is 3. Using this result of Silhouette measure, we performed K-means clustering with K=3. From the cluster analysis, we found that the size of cluster 1, 2, and 3 were 925, 683, and 2598, respectively. Next we carried out PCA and plotting according to three technological clusters. The left plot in Fig. 1 shows the PCA plot of LED technology cluster 1. In Fig. 1, each number represents an individual object (patent document), and each term represents a technology keyword. And, the larger the length of the arrows displayed with each keyword, the better represent the technology representation. We found that the keywords of ‘control’, ‘lamp’, ‘modul’, circuit’, and ‘wireless’ are major keywords representing the sub-technology of cluster 1. Next, we show the PCA plot of LED technology cluster 2 in center of Fig. 1, 2, and 3.

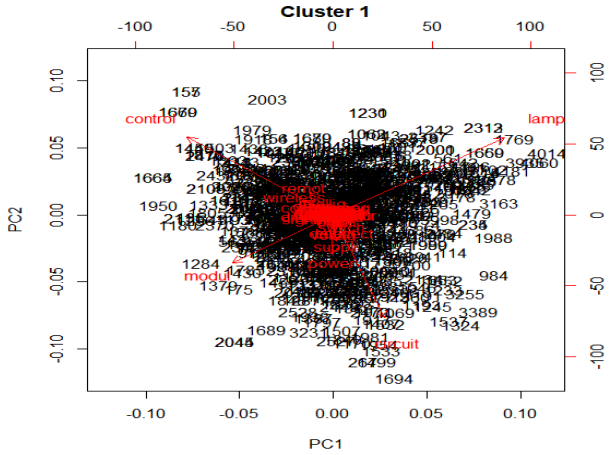

Fig. 1: PCA Plot of LED Technology by Cluster 1.

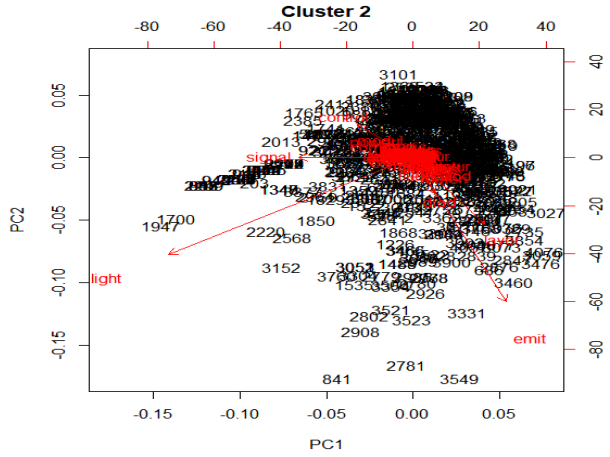

Fig. 2: PCA Plot of LED Technology by Cluster 2.

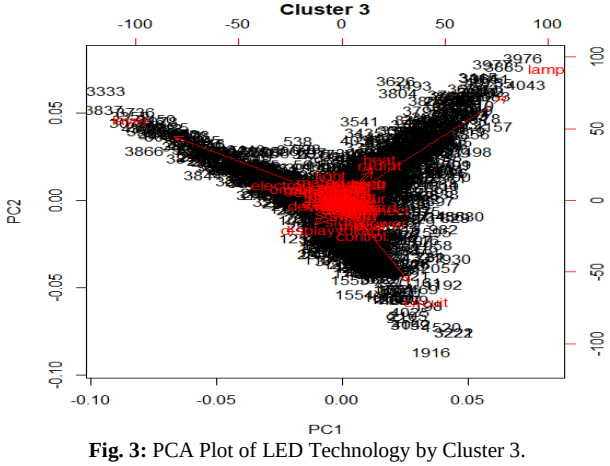

Fig. 3: PCA Plot of LED Technology by Cluster 3.

The sub-technology cluster 2 of LED is based on the keywords of 'control', 'signal', 'light', 'emit', and 'layer'. In Fig. 1, the 'control' keyword belongs to technology clusters 1 and 2 at the same time. This implies that this keyword-based technology links the sub-technologies of clusters 1 and 2. The technology cluster 3 is the largest cluster in the LED technology. The PCA plot of LED technology cluster 3 is represented in right of Fig. 1. This technology cluster is led by the 'layer', 'lamp', and 'circuit' keywords, and the 'organ' and 'circuit' keywords can be added in Fig. 1. The 'lamp' and 'circuit' keywords are common keywords of clusters 1 and 3, in addition, the 'layer' keyword is included in clusters 2 and 3 at the same time. Table 1 shows the results of computing principal component loading values to confirm specific representations of the main keywords identified through the three PCA plots of LED technology. In Table 1, we represent the top five keywords with highest loading on absolute values.

Table 1: PCA Results by Technology Clusters

Technology clusters	Representative keywords (loading value)
Technology 1	circuit (0.1976), control (-0.5619), lamp (0.6575), modul (-0.3880), wireless (-0.1196)
Technology 2	control (-0.1789), emit (0.2986), layer (0.2297), light (-0.7838), signal (-0.3666)
Technology 3	circuit (0.2483), electrod (-0.1785), lamp (0.6036), layer (-0.6208), organ (-0.1504)

Using the results of PCA plots and loading values, we built the technological structure of LED technology in Fig. 4.

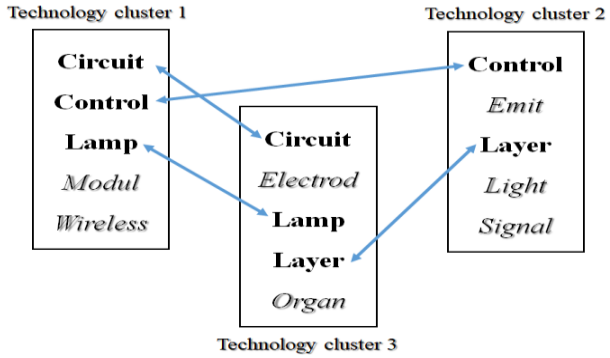

Fig. 4: Technological Structure of LED Technology.

Technology cluster 1 and 3 are connected each other via the technologies based on the 'circuit' and 'control' keywords. In addition, the technology of 'layer' connects the technology clusters 2 and 3. Lastly technology clusters 1 and 2 are connected by the 'control' keyword-based technology. In this LED technology structure, technology cluster 1 is based on 'modul' and 'wireless'. The technology by the keywords of 'emit', light', and 'signal' represents the technology cluster 2. Also, the 'electrod' and 'organ' keywords are the basis of technology cluster 3. In this paper, we knew the LED technology consists of three sub-technologies are organically connected to each other through various technology keywords.

4. Conclusion

In this paper, we proposed an analytical method of scientific and industrial keyword data using structured covariance and clustering. We used the technological keywords extracted from patent documents for scientific and industrial keywords. In addition, we considered PCA and K-means clustering for structured covariance and clustering. To extract technology keywords from patent documents, we used various text mining techniques. Using the Silhouette width, we determined the optimal number of clusters for K-means clustering. The number of determined clusters represented the number of the sub-technologies for target technology. In our research, the target technology was LED, and the number of sub-technologies was three. The technology structure was constructed by using the results of PCA plots and loading values for each technology cluster. Using the result of technology structure of LED in our experiments, the companies related to LED technology can make the technology development policy to improve their technological competitiveness in the market. This research contributes to diverse fields of technology management such as R&D planning, new product development, and technology forecasting. At present, our technology analysis uses principal PCA and clustering. However, using a more advanced statistical analysis model will improve the performance of the prediction for technology keyword analysis. This part remains one of our future research works.

References

[1] N. Rezki, O. Kazar, L. H. Mouss, L. Kahloul and D. Rezki, "A Hybrid Approach for Complex Industrial Process Monitoring", *Journal of Scientific and Industrial Research*, 76, 608-613, 2017.
[2] S. Jun, S. Park and D. Jang, *Patent Analysis and Technology Forecasting*, Seoul, Korea, Kyowoo, 2014.
[3] S. Jun, "Technology analysis of artificial intelligence using Bayesian inference for neural networks", *International Journal of Engineering & Technology*, 7(2.3), 43-45, 2018.
[4] J. Kim and S. Jun, "Graphical Causal Inference and Copula Regression Model for Apple Keywords by Text Mining", *Advanced Engineering Informatics*, 29(4), 918-929, 2015. https://doi.org/10.1016/j.aei.2015.10.001.
[5] D. Uhm, J. Ryu and S. Jun, "An Interval Estimation Method of Patent Keyword Data for Sustainable Technology Forecasting", Sustainability, 9(11), 2025, 2017. https://doi.org/10.3390/su9112025.
[6] S. Hicks, *Patent Mining Searches, A Patent Searcher's Quick Reference*, Asheville, NC, Wolf Mountain IP, 2017.
[7] I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning*, Cambridge, MA, MIT Press, 2016.
[8] R. E. Schumacker, *Using R with Multivariate Statistics*, Singapore, SAGE, 2015.
[9] E. Alpaydin, *Machine Learning*, Cambridge, MA, MIT Press, 2017.
[10] S. Jun, S. Park and D. Jang, "Document Clustering Method Using Dimension Reduction and Support Vector Clustering to Overcome Sparseness", *Expert Systems with Applications*, 41(7), 3204-3212, 2014. https://doi.org/10.1016/j.eswa.2013.11.018.
[11] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, C. Chang and C. Lin, *Package 'e1071'*, CRAN R Project, 2017.
[12] A. Kassambara, *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning*, STHDA, 2017.
[13] R Development Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org, 2017.
[14] I. Feinerer and K. Hornik, *Package 'tm' Ver. 0.7-3, Text Mining Package*, CRAN of R project, 2017.
[15] USPTO, *The United States Patent and Trademark Office*, http://www.uspto.gov, 2017.
[16] WIPSON, *WIPS Corporation*, http://www.wipson.com, 2017.