ICIC Express Letters
Part B: Applications
Volume **12**, Number **12**, December **2021**
 ISSN 2185-2766
pp. **1161–1168**

KEYWORD ANALYSIS IN AGRICULTURE-FOOD SECTOR OF KOREA'S SCIENCE AND TECHNOLOGY INFORMATION SERVICE

Bom Yun[1], Ji-Youn Jeong[2], Joonsoo Bae[1,*] and Jong-Il Yoon[3]

[1]Department of Industrial and Information Systems Engineering
Jeonbuk National University
567, Baekje-daero, Deokjin-gu, Jeonju-si, Jeollabuk-do 54896, Korea
yspring@jbnu.ac.kr; *Corresponding author: jsbae@jbnu.ac.kr

[2]Food Standard Research Center
Korea Food Research Institute
245, Nongsaengmyeong-ro, Iseo-myeon, Wanju_Gun, Jeollabuk-do 55365, Korea
allenjeong@kfri.re.kr

[3]Industrial Innovation Strategy Lab.
Korea Construction Equipment Technology Institute
36, Sandan-ro, Gunsan-si, Jeollabuk-do 54004, Korea
jiyoon@koceti.re.kr

Received April 2021; accepted June 2021

Abstract. *In order to investigate the trend of Korea's science and technology policy in the field of agriculture-food sector, it is assumed that a keyword is included in the project titles that has been selected as a national R&D program and received government funding. This is proven through text mining, CONCOR analysis, and regression analysis. First of all, the project title data were collected through National Science & Technology Information Service system (NTIS system). Through the analysis, seven keywords were extracted. With N-gram analysis, it was found that these keywords have characteristics that are connected to each other. It was divided into 8 groups with similar meanings through CONCOR analysis. Among them, the 'food industry' group contains the most frequency. Through regression analysis, it was proved that there is a proportional relationship between the keyword frequency, government funds, and paper performance. As a result, the trend of Korea's national science and technology policy in the field of agriculture-food is 'food industry'. In conclusion, project titles that contain high-frequency keywords can receive a lot of government funds, and it is proportional to the paper performance.*
Keywords: Agriculture-food, Policy trend, Text mining, CONCOR analysis, Regression analysis

1. **Introduction.** In 2021, Korean government invested 27.4 trillion-won in national R&D programs in each ministry. Among them, the agriculture-food sector accounts for about 10% (275.7 billion-won) [1]. In particular, agriculture-food R&D has a unique mechanism called a guidance system. This is the process by which research results are transmitted to the field, but not in other fields. And there is a point of view that this field must respond to changes in the market environment and the natural environment. Therefore, the speed of reflecting new technologies is different from other fields. Therefore, it is necessary to identify individual research trends [2].

In recent research, text mining and network analysis are being used to confirm the trend in various fields. Its fields include bio-health, music therapy, cosmetics, and wearables [3-6]. In past research, research achievements such as patents were used to identify research trends in science and technology. And there are studies that conducted analysis based

DOI: 10.24507/icicelb.12.12.1161

on keywords described in the summary. Representatively, Park et al. performed keyword network analysis using US patent abstracts. And based on this, it is predicted to science and technology trends [7]. Kim calculated the frequency based on the English keywords in the summary. He constructed a research trend network using the frequency of research keywords for each detailed technology. And the policy trends were identified by performing network analysis [8].

It was conducted only to keyword analysis research on agricultural R&D by Kim and Kim [2]. In this paper, text mining and TF-IDF were calculated based on the Korean keywords in the abstract. And trend analysis was performed by grouping by similar keywords. At this time, the agricultural R&D list was used to group keywords.

In this study, research trends are identified not only in the agricultural field but also in the food field. And it is reflected of the characteristics of agricultural R&D. Therefore, trends can be identified more quickly than existing research cases. For this purpose, it is analyzed that the project title was selected as national R&D program in the agriculture-food sector. Text mining is the process of extracting implicit information between text data [9]. In this study, text mining was performed using Textome.

Also, keywords acquired through text mining are grouped. For this, the CONCOR analysis was used, not the agricultural R&D list. CONCOR analysis is a part of network analysis method. A network is one of the expression methods representing various types of systems created by humans or things. Keyword network analysis uses this principle. Therefore, this method analyzes the co-occurrence relationship of keywords in the text and infers the meaning. CONCOR analysis is one of the network analysis methods. In this way, nodes can be identified and relationships can be identified based on Pearson's correlation analysis on co-occurring keywords [10]. In this study, CONCOR analysis is performed using UCINET6. It is possible to derive a cluster of keywords with similarities. And the relationship between keywords was visualized through NetDraw.

Finally, regression analysis was performed. This method is performed to understand the relationship between keyword frequency, government funding, and thesis performance. In this case, Minitab 17 was used. Based on the results, it is identified of the relationship between 7 major keywords and national R&D program. And it will predict the trend of Korea's science and technology policy in the field of agriculture-food.

The rest of this paper is organized as follows. In Section 2, we introduce data analysis method and procedures. Section 3 is data result; we describe the results in detail. Finally, we conclude this paper in Section 4.

2. Data Analysis Method and Procedures.

2.1. Data collection and text mining.

For the research, data on project title, government funds, and paper performance are collected. Data are collected by searching "Ministry of Agriculture, Food and Rural Affairs" through National Science & Technology Information Service system (NTIS system). This data is about the project titles, government funds, and paper performance from 2008 to June 2021. Duplicate data for all columns among the acquired data is deleted. There remain the 16,987 data of the project title-government funds and 11,462 data of the paper performance.

Data filtering is performed to increase the accuracy. In this case, Python's KoNLPy package is used [11]. And there obtain the results of the data about the top thousands of keyword frequency. At this time, stop words are eliminated from the data of the project title. They are "development", "technology", "research", "use", "system", etc. After that, text mining is performed through Textome and the top 50 keyword frequencies are selected as shown in Table 1. This is a cumulative ratio of about 25%.

The analysis results also included TF-IDF and degree centrality. TF-IDF stands for "Term Frequency-Inverse Document Frequency". This is a method of evaluating the

Table 1. The result of keyword frequency (top 50)

No.	Keyword	Frequency	Rate	Cumulative Rate	No.	Keyword	Frequency	Rate	Cumulative Rate
1	function	1834	1.48662121961%	1.48662121961%	26	plant	492	0.398810054553%	17.6465343244%
2	production	1721	1.39502460139%	2.88164582101%	27	Control	484	0.392325338218%	18.0388596626%
3	food	1661	1.34638922889%	4.22803504989%	28	characteristic	480	0.389082980051%	18.4279426427%
4	material	1475	1.19561957412%	5.42365462401%	29	Agriculture	473	0.383408853259%	18.8113514959%
5	exportation	1238	1.00350985272%	6.42716447672%	30	cultivate	461	0.373681778758%	19.1850332747%
6	product	1236	1.00188867363%	7.42905315036%	31	optimal	456	0.369628831049%	19.5546621057%
7	industry	1132	0.917587361288%	8.34664051164%	32	safety	449	0.363954704256%	19.91861681%
8	Model	865	0.701159953634%	9.04780046528%	33	Environment	441	0.357469987922%	20.2760867979%
9	animal	755	0.611995104039%	9.65979556932%	34	mass	432	0.350174682046%	20.62626148%
10	variety	724	0.586866828244%	10.2466623976%	35	crops	428	0.346932323879%	20.9731938039%
11	virus	700	0.567412679242%	10.8140750768%	36	disease	408	0.330720533044%	21.3039143369%
12	health	668	0.541473813905%	11.3555488907%	37	facility	398	0.322614637626%	21.6265289745%
13	domestic	657	0.532557328945%	11.8881062197%	38	develop	394	0.319372279459%	21.945901254%
14	Diagnosis	647	0.524451433528%	12.4125576532%	39	quality	393	0.318561689917%	22.2644629439%
15	ground	604	0.489596083231%	12.9021537364%	40	extract	389	0.31531933175%	22.5797822756%
16	vaccine	594	0.481490187814%	13.3836439242%	41	prevention	386	0.312887563125%	22.8926698388%
17	microbe	577	0.467710165603%	13.8513540898%	42	activation	383	0.310455794499%	23.2031256333%
18	efficacy	567	0.459604270186%	14.310958836%	43	resource	380	0.308024025874%	23.5111496591%
19	pig	538	0.436097173474%	14.7470555335%	44	Produce	376	0.304781667707%	23.8159313269%
20	Standard	530	0.42961245714%	15.1766679906%	45	matter	370	0.299918130456%	24.1158494573%
21	smart	530	0.42961245714%	15.6062804478%	46	feed	369	0.299107540915%	24.4149569982%
22	Fermentation	512	0.415021845388%	16.0213022932%	47	Bio	365	0.295865182747%	24.710822181%
23	origin	508	0.411779487221%	16.4330817804%	48	ingredient	362	0.293433414122%	25.0042555951%
24	processing	508	0.411779487221%	16.8448612676%	49	foot-and-mouth disease	362	0.293433414122%	25.2976890092%
25	manufacture	497	0.402863002262%	17.2477242699%	50	natural	362	0.293433414122%	25.5911224233%

Table 2. The results of TF-IDF and degree centrality (partial)

Keyword	Frequency	RANK	TF-IDF	RANK	Degree centrality	RANK
function	1834	1	4210.875228	1	0.069308545	2
production	1721	2	4033.98433	2	0.076810176	1
food	1661	3	3920.086848	3	0.047455969	6
material	1475	4	3640.412038	4	0.056099152	4
exportation	1238	5	3358.270217	5	0.057566862	3
product	1236	6	3268.008867	6	0.049086758	5
industry	1132	7	3093.064653	7	0.045172864	7
Model	865	8	2589.678647	8	0.035388128	10
animal	755	9	2383.579815	9	0.03359426	11
variety	724	10	2300.910973	10	0.033268102	13
virus	700	11	2275.980334	11	0.037997391	8
health	668	12	2217.151831	12	0.027723418	30
domestic	657	13	2146.246999	14	0.032289628	16
Diagnosis	647	14	2158.075163	13	0.03196347	17
ground	604	15	2019.66903	16	0.035714286	9

importance of a word in a text. If this value is higher, it means that it is the main keyword in the text [12].

Degree centrality is a method of measuring how many connections between keywords. If this value is higher, it means that there are many linked keywords. Through this, it can be explained to the main keywords on the network. And it is possible to describe the distance to the main keywords [7]. As shown in Table 2, TF-IDF and degree centrality generally tend to increase together as the higher the keyword frequency.

2.2. **N-gram analysis.** N-gram analysis is a method of identifying the relationship between keywords through a network graph. In this way, it is possible to confirm how many sub-structures are formed in the network. This means that when keyword 1 appears, keyword 2 appears at the same time [12]. In addition, if the frequency of keyword 1 and keyword 2 is high, it means that the number of times the two words appear simultaneously is high. In this study, Table 3 shows the N-gram analysis results. Figure 1 is a visualization based on the top 50 N-grams.

2.3. **CONCOR analysis.** CONCOR analysis is performed to derive clusters based on the top 50 keyword frequencies. As a result of the analysis, they are divided into 8 groups as shown in Figure 2. The groups are 'Crop Cultivation', 'Plant/Bio', 'Virus', 'Hog-raising', 'Food Industry', 'Microorganism', 'Livestock', and 'Process/Standardization'.

TABLE 3. The result of N-gram analysis (partial)

Word 1	Word 2	Frequency
function	food	438
health	function	380
mass	production	287
higher	value	245
food	material	227
variety	develop	205
smart	farm	203
avian	influenza	179
function	material	144
Diagnosis	law	128
material	product	128
pet	animal	119
foot-and-mouth disease	vaccine	113
production	processing	111
pig	fever	109
feed	additive	109
Produce	processing	106

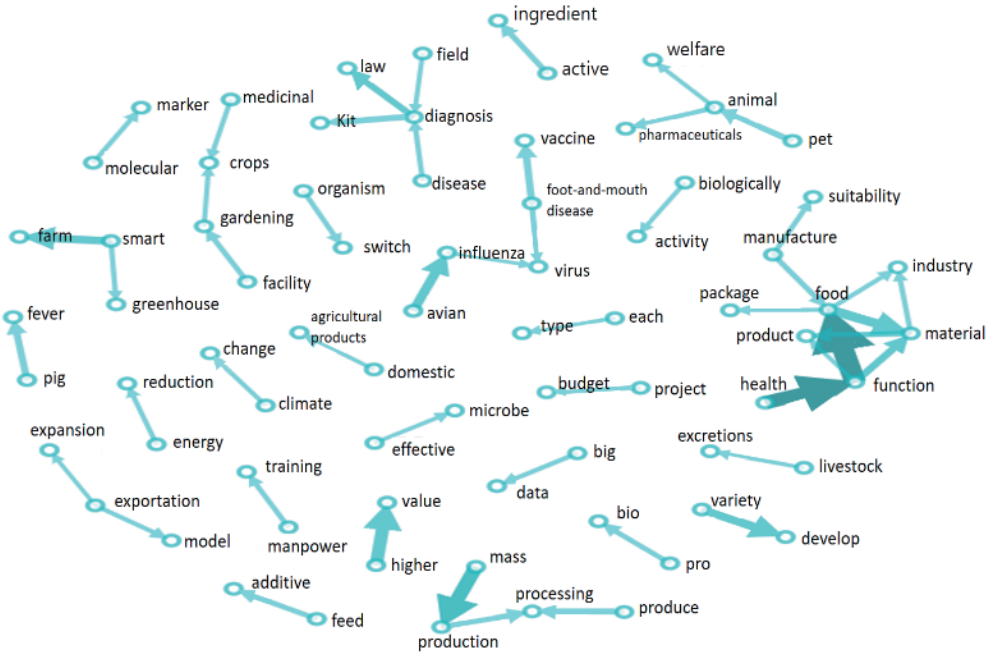

FIGURE 1. A visualization of N-gram network (top 50)

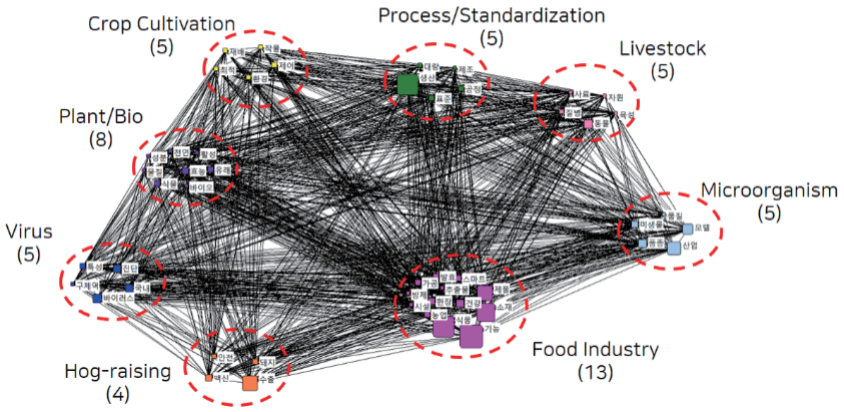

FIGURE 2. The result of CONCOR analysis

2.4. **Regression analysis.** Regression analysis is performed to confirm the relationship between keyword frequency, government funds, and paper performance. According to the top 50 keyword frequencies, they are summarized of the sum and average of government funds and paper results in Table 4. The sum and average are calculated based on the project title including keywords.

TABLE 4. Based on the keyword frequencies, the sum and average of government funds and paper performance (partial)

Keyword		Frequency	Funds (million)		Paper	
			Sum	Average	Sum	Average
1	function	1,859	169,871	93	1,247	3
2	production	1,651	170,189	104	1,094	4
3	food	1,638	235,939	145	1,203	4
4	material	1,483	145,451	99	1,018	4
5	exportation	1,145	122,234	107	628	4
6	product	1,351	109,684	84	805	3
7	industry	1,130	119,922	108	758	4
8	Model	895	109,767	123	453	4
9	animal	739	65,985	89	510	4
10	variety	715	68,552	96	297	3
11	virus	688	58,671	85	603	3
12	health	622	61,555	101	360	3
13	domestic	708	65,165	92	522	3
14	Diagnosis	631	57,065	91	472	3
15	ground	599	56,392	96	306	4

Regression analysis is performed using Minitab. The Pearson correlation coefficient (R^2) ranges from 0.1 to 1.0. The results can be interpreted as follows.

- If $R^2 = 1.0$, then perfect linear correlation.
- If $0.7 \leq R^2 < 1.0$, then the high linear correlation.
- If $0.4 \leq R^2 < 0.7$, then the median linear correlation.
- If $0.1 \leq R^2 < 0.4$, then the low linear correlation.

It is 0.867 of the Pearson correlation coefficient between the keyword frequencies and government funds. And it is 0.872 of the Pearson correlation coefficient between keyword frequency and thesis performance. Thus, there is a high linear correlation between the keyword frequencies, government funding, and paper performance.

As a result, the relationship between keyword frequency and government funding is explained by Equation (1). And the relationship between keyword frequency and paper performance is explained by Equation (2).

$$y = 109.1 + 0.008170F \tag{1}$$

$$y = 84.13 + 1.315P \tag{2}$$

(F: Government funds, P: Paper performance)

This can be respectively represented by increasing graphs as shown in Figure 3 and Figure 4.

3. Result Analysis.

3.1. **The top 7 keywords selection.** In Table 2, there are similarity to the keyword frequency ranking and the TF-IDF ranking. However, the ranking of the degree centrality tends to be a bit different except for the top 7 keywords. Therefore, the top seven keyword frequencies are selected as main keywords. They are 'function', 'production', 'food', 'material', 'exportation', 'product', 'industry'. Table 5 shows the results of text mining of main keywords.

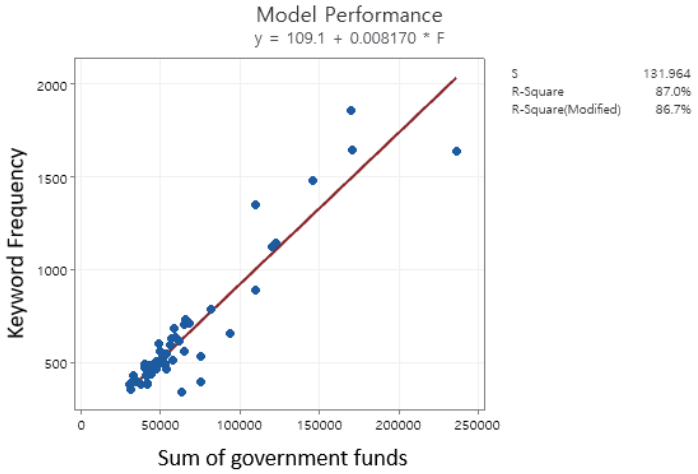

FIGURE 3. The result of the regression analysis between the keyword frequency and government funds

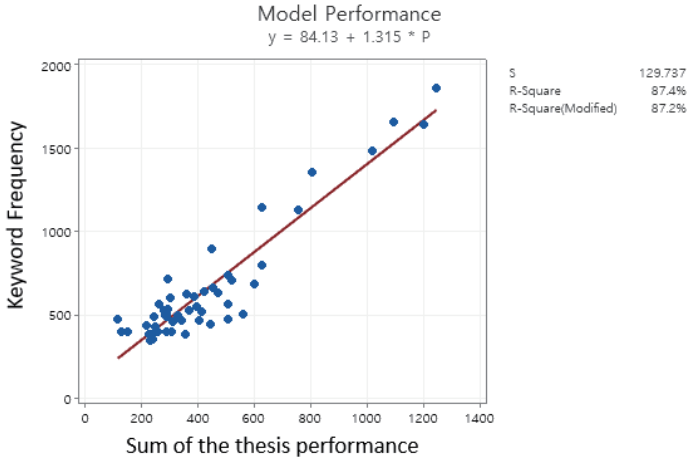

FIGURE 4. The result of the regression analysis between the keyword frequency and paper performance

TABLE 5. The results of text mining of main keywords

Keyword	Frequency	RANK	TF-IDF	RANK	Degree centrality	RANK
function	1834	1	4210.875228	1	0.069308545	2
production	1721	2	4033.98433	2	0.076810176	1
food	1661	3	3920.086848	3	0.047455969	6
material	1475	4	3640.412038	4	0.056099152	4
exportation	1238	5	3358.270217	5	0.057566862	3
product	1236	6	3268.008867	6	0.049086758	5
industry	1132	7	3093.064653	7	0.045172864	7

3.2. **Identifying connections between keywords.** In Figure 1, there is a group that forms the most relationship in N-gram analysis. This group is composed of 10 keywords ('health', 'function', 'food', 'agriculture', 'packaging', 'industry', 'material', 'product', 'manufacture', 'characteristic'). There is one of them that the five keywords ('function', 'food', 'industry', 'material', 'product') correspond to the main keywords. Also these keywords had featured linked each other as shown in Figure 5.

3.3. **Main group.** The 50 keywords are classified into 8 groups through CONCOR analysis. The main keywords are distributed in 4 groups out of 8 groups. Four of the main keywords are included in 'food industry' group. These keywords are 'function', 'food', 'material', and 'product'. The sum of the frequency of the main keyword is 6206 times.

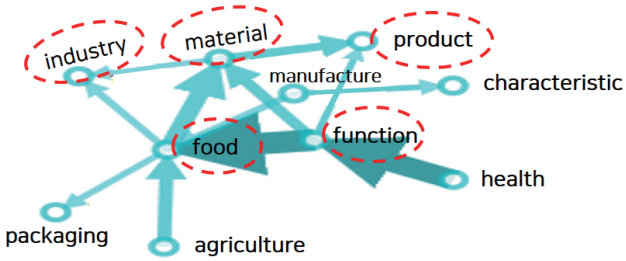

FIGURE 5. The group with the most relationship in N-gram analysis

TABLE 6. The top 5 keywords and frequency by group through CONCOR analysis

Group	The top 5 keyword frequencies
Food Industry	function(1834), food(1661), material(1475), product(1236), health(668)
Process/standardization	production(1721), standard(530), processing(508), mass(432), produce(376)
Microorganism	industry(1132), model(865), variety(724), microbe(577), quality(393)
Hog-raising	exportation(1238), vaccine(594), pig(538), safety(449)
Livestock	animal(755), disease(408), develop(394), resource(380), feed(369)
Virus	virus(700), domestic(657), diagnosis(647), characteristic(480), food-and-mouth disease(362)
Plant/Bio	efficacy(567), origin(508), plant(492), activation(383), matter(370)
Crop Cultivation	control(484), cultivate(461), optimal(456), environment(441), crops(428)

One main keyword is included in each of the 'process/standardization', 'microorganism', and 'Hog-raising' groups. The main keywords are concentrated in the 'food industry' group. Based on this, it can be seen that national R&D program are focused on the 'food industry' group.

3.4. **The relationship between keyword frequency, government funds, and paper performance.** There is proportional about keyword frequency – government funding, keyword frequency – paper performance as Equations (1) and (2). As mentioned in Section 2.4, it is 0.867 of the Pearson correlation coefficient between the keyword frequencies and government funds. And it is 0.872 of the Pearson correlation coefficient between keyword frequency and paper performance. Thus, there is a high linear correlation.

Therefore, a lot of government funds are invested in the project titles that contain words with high keyword frequency. And a lot of paper results are created with this high keyword frequency.

4. **Conclusions.** This study is to predict policy trends through the project titles that received government funds in agriculture-food sector of South Korea. For this purpose, it is assumed that keywords are included in the project title. Data are collected through National Science & Technology Information Service system (NTIS system). This data is about the project titles, government funds, and paper performance from 2008 to June 2021 in "Ministry of Agriculture, Food and Rural Affairs". Trend is analyzed through text mining, keyword network analysis, and regression analysis.

The results can be summarized as follows. First, using data mining techniques such as N-gram analysis and keyword frequency, TF-IDF, and connection centrality, the main keywords are extracted, which are top 7 keywords, "function", "production", "food", "material", "export", "product", and "industry". Second, using regression analysis, keyword frequency vs. government funds and keyword frequency vs. thesis performance are found to be strongly related. Currently, the Pearson correlation coefficients are 0.867 and 0.872, respectively. When the Pearson correlation coefficient is 0.7 or higher, it is interpreted as "high linear correlation". That is there is close relationship between keyword frequency

and government funding, also between keyword frequency and thesis performance. Third, as the result of CONCOR analysis, the main keywords are included most in the 'food industry' group. Eight groups were divided into CONCOR analysis. Among them, 4 of the top 7 keywords are included in the 'food industry' group. The four keywords are "function", "food", "material", and "product". These keywords are ranked 1st, 3rd, 4th, and 6th respectively. Therefore, Korean government invests the most funds in the 'food industry' group in the agriculture-food sector. Also 'food industry' group produces most thesis outputs.

This study has the following significance. Research trends in agriculture-food sector are analyzed based on the project titles funded by Korean government. So the current hot trends can be said 'food industry'. We prove that Korean government funds are focused on 'food industry' group and the most thesis performance are created in this group.

However, this study has three limitations as follows. First, the used data for the analysis were 14 years period from 2008 to June 2021, and this data is analyzed at once. For this reason, it is possible that the selected main keywords may not reflect the trend accurately. Second, since each research institute has a different method of registering and managing performance data, there may be overlapping outputs. In addition, they are likely to be missing outcomes. Lastly, Korea's national R&D programs are contracted on a one-year basis, so some of the project titles may be duplicated. Therefore, it is considered that more reliable analysis results can be obtained if a follow-up study is conducted.

Acknowledgment. This work is partially supported by the Spatial Information Research Institute grant funded by LX (Grant No. 2020-502).

REFERENCES

[1] *2021 Government R&D Project Online Ministries Joint Briefing Session – Main Features of the Government R&D Budget in 2021*, Ministry of Science and Technology Information and Communication, Republic of Korea, 2021.

[2] J. H. Kim and S. S. Kim, A study on the analysis of agricultural R&D keywords using textmining method, *Journal of the Korea Academia-Industrial Cooperation Society*, vol.22, no.2, pp.721-732, 2021.

[3] S. H. Park, Y. M. Yun, H. Y. Kim and J. S. Kim, Technology convergence & trend analysis of biohealth industry in 5 countries: Using patent co-classification analysis and text mining, *Journal of the Korea Convergence Society*, vol.12, no.4, pp.9-21, 2021.

[4] E. Y. Hwang, Analysis on topic modeling and trend of "Korean Journal of Music Therapy" using text mining (1999-2019), *Korean Music Therapy Association*, vol.22, no.2, pp.29-47, 2020.

[5] H. S. Lim and J. W. Shin, A study on the trends of cosmetics through big data analysis – Focusing on text mining and sematic network analysis, *Journal of the Korean Society of Illustration Research*, vol.66, pp.85-95, 2021.

[6] M. J. Kim, Analyzing the trend of wearable keywords using text-mining methodology, *Journal of Digital Convergence*, vol.18, no.9, pp.181-190, 2020.

[7] J. S. Park, N. R. Kim and E. J. Han, Analysis of trends in science and technology using keyword network analysis, *Journal of the Korea Industrial Information Systems Research*, vol.23, no.2, pp.63-73, 2018.

[8] J. H. Kim, *A Study on Research Trends of National R&D in IT Sector Using Keyword Network Analysis*, Soongsil University, 2014.

[9] T. H. Cho, The concepts and applications of text mining, *Journal of Scientific & Technological Knowledge Infrastructure*, Republic of Korea, vol.5, pp.76-85, 2001.

[10] W. G. Kang, E. S. Ko, H. R. Lee and J. N. Kim, A study of the consumer major perception of packaging using big data analysis –Focusing on text mining and semantic network analysis–, *Journal of the Korea Convergence Society*, vol.9, no.4, pp.15-22, 2018.

[11] E. J. Lee and S. Z. Cho, KoNLPy: Korean natural language processing in python, *Korean Language and Korean Information Processing Conference*, pp.133-136, 2014.

[12] J. Y. Jung, Keyword analysis of performing arts selection criteria using TF-IDF and N-gram: Focused on projects to supports state subsidies for public institutions, *Korea Humanities Content Society*, vol.58, pp.253-282, 2020.