International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 11 (2019) pp. 001-016

Received: 30 July, 2018; Accepted: 25 January, 2019; Publish: 12 February, 2019

Academic Conference Analysis for Understanding Country-Level Research Topics Using Text Mining

Kwanho Kim[1], Sue-Kyung Lee[1], Heemin Park[2] and Jinseok Chae[3*]

[1] Department of Industrial Management and Engineering, Incheon National University,
Incheon 22012, South Korea
khokim@inu.ac.kr, dltnrud1212@naver.com

[2] Department of Software, Sangmyung University,
Cheonan 31066, South Korea
heemin@smu.ac.kr

[3] Department of Computer Science and Engineering, Incheon National University,
Incheon 22012, South Korea
jschae@inu.ac.kr

***Abstract*: The importance of academic conferences is getting intensively larger as a way to publish the up-to-date research results on each particular research topics in a fast manner unlike journals, and the number of conferences tends to be increased year by year. Moreover, since a conference information, mostly accessible on the Internet, contains not only topics but also geographical areas where the conference was held, these are considered as a valuable source to understand the research trends according to countries. In this paper, we aim to develop methods for analyzing country-level research trends and the relationships among the countries by using text mining and clustering techniques. Specifically, we collected conference information from 8,957 websites from 2015 to 2017, and we found three clusters of countries according to their distributions of topics and the similarities among them. The experimental results show that some countries focus on various topics ranging from social science and medicine, while the others mainly concentrated on some particular topics such as engineering. Moreover, we found country groups that show quite similar in terms of topics. For instance, the following three country groups are found (Philippines, Indonesia, Thailand), (China, Japan, Hong Kong), and (Austria, Czech Republic, Netherlands).**

***Keywords*:** Academic conference analysis, Big data analysis, Data mining, Text mining, Topic analysis, Country clustering.

I. Introduction

Recently, as academic conferences play an essential role to publish the state-of-the-art research results, these are considered as an important source to capture the current research trends in a timely manner [1, 2]. Moreover, the conference information is able to be utilized to understand how much a country focuses on a particular topic since more conferences in a topic likely held in a country as more interests on that topic in the country.

The increasing number of academic conferences seems natural mainly due to the in-depth developments of each individual research areas, resulting in the needs of more conferences that cover particular and sophisticated academic areas. As there is no central organization that monitors the statuses of academic conferences, the exact numbers of conferences around the world still remain as an unknown. Based on our database collected for this research, the numbers of conferences had significantly increased over the past several years. In particular, it was about 3,000 in 2010, while there were over 5,000 conferences around the world in 2016.

The conference information is usually maintained or developed by various collaborations and partnerships through the publication and dissemination of many research results of researchers from various countries. In addition, based on the time and geographical location information of the conferences, current research topics that are of interest to the community and the country can be identified [3]. Furthermore, conference information can be used as valuable data to figure out which research topics are trending in which countries [4]. In the case of journal papers, it takes one to two years from the start of the research to the completion of the final publications. On the contrary, since the research topics covered in conferences are usually very sensitive to the current trends not only for research areas but also for industrial sectors, it is considered that the most trending research can be found on the related conference.

Therefore, this research aims to find the distributions of research topics according to country and discover the hidden relationships among countries in terms of their research trends. Specifically, we propose a method that can be used for figuring out the research trend by extracting topic information that can clearly identify the research topic from a large amount of academic information. Our method attempts to identify topical concentration levels and topical diversification levels of all countries on research topics. For this purpose, the conference information is represented by topic vectors for multiple research topics through Latent Dirichlet Allocation (LDA) topic modeling. The expertise and diversity can be

identified by the standard deviation of each country's topic vector. Therefore, if the standard deviation is high, it is determined to be highly concentrated because it is biased toward a specific topic. If it is low, it can be determined that diversity is high.

Additionally, country clustering based on research themes based on research similarity networks of all countries is conducted. This enables us to determine the relevance of research between countries, which provides useful results on common interests and research collaboration opportunities. The network is constructed through modularity analysis, and the main topics of each cluster and identify the research topics that are common to the countries in each cluster are discovered.

Finally, the changes in international major researches over time are measured and the countries that lead each research topic are identified. To this end, we calculate the average of topic interest levels for each year. The time series graph shows how each theme changes over time, so the trends of interest in a topic can be seed. In addition, based on the top five countries, we can identify the leading cities for those topics.

II. Related Work

There have been some research results that aims to analyze academic conference information. Most of them have focused on conference topic understanding by using published papers, abstractions, and references. Jin et al. [2] reported the research and development trends in South Korea by using the research networks constructed through data about human resources and research projects from national project databases. By grouping the research and development organizations and experts in the obtained network, the current status of convergence technology and the important experts that lead the technology convergence in South Korea were discovered. More recently, Muhammad et al. [5] proposed another new way to understand the contribution of researchers' articles originate primarily from China, USA, and European countries in a particular academic field which is the organic Rankine cycle using an academic journal database. Similarity, Zhuang et al. [6] suggested heuristic methods to automatically discover the quality of academic conferences by mining the characteristics of the program committee members involved in each conference.

Topic and trend analysis has been explored in patent documents using text mining as well. Tseng et al. [7] analyzed patent document for patent classification, text segmentation, summary extraction, and topic identification. Tang et al. [8] proposed a topic-driven patent analysis and mining system for the patent network which used probabilistic model to characterize the topical evolution.

Topic analysis can be used for paper recommendation system. Pan and Li [9] adopted the collaborative filtering techniques and proposed a paper recommendation system for researchers based on topic modeling techniques using the text of papers.

To find hot and trending topics from science research papers, Griffiths and Steyvers [10] used Markov chain Monte Carlo algorithm and Bayesian model selection technique to extract topics. Other attempted has been done for discovering hot topic trends by Rajaraman and Tan [11]. They proposed a topic detection and tracking system with clustering technique using neural networks. Similar with our approach, Wang and McCallum [12] adopted LDA-style topic model and analyzed trending topics over time.

Although there have been research on topic modeling and analysis on research papers and patents, there has been no research on country-level analysis of research topics. To the best of our knowledge, our research is the first attempt for country-level research topic analysis.

III. Proposed methods

In this section, the proposed methods of research trend analysis are divided into three processes. The first is to analyze the degree of concentration on research by country. Secondly, we perform national clustering based on the research similarity network. Finally, we perform a time series analysis of the research subject according to geographical information.

A. Topic vector generation using topic modeling

Since a specific society is likely to be associated with many research topics due to the multidisciplinary nature of modern academic societies, it is difficult to limit the research topics on a particular conference or society. In order to take this into account, a conference information is represented as a topic vector based on the relevance scores of all research topics for each conference document.

The topic vector generation process uses the LDA algorithm which is one of the popular methods for topic modeling techniques. The LDA algorithm is a stochastic model for the inference of the ratio of a particular topic in a document. It stochastically probes the appearance frequency of words in one text and learns the ratio of a topic [13]. The LDA deduces the probability distributions of the topics that the document can have and the probability distributions of the words for each topic [14]. In other words, *K* topics from the entire document set can be extracted, and each document has relevance scores for the topics.

Specifically, the topic vector of particular conference d_n among *N* conferences is defined as the following

$$\boldsymbol{d_n} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_k}, \cdots, \boldsymbol{x_K}\} \quad (1)$$

where x_k is the *k*-th topic relevance score of d_n.

In this study, the number of selected topics, denoted as *K*, is empirically defined as 31, and Table 1 shows the keyword list of selected topics.

B. International research trend analysis

The research concentration analysis attempts to investigate the relative concentration of research topics in each country. This can identify countries with diversity and expertise in research topics. To analyze the research concentration of countries, a country's topic vector is generated based on the topic vector of each conference document. The topic vector of country c, denoted as V_c, is expressed as $\{v_1, v_2, \cdots, v_k, \cdots, v_K\}$. We note that the dimension of the topic vector of a conference is the same for that of a country since the topics for both conference and country are represented as *K* topic elements. Here, the *k*-th topic elements for a country,

denoted as v_k, is defined as a ratio of conferences related to the topic, and it is specifically defined as

$$\boldsymbol{v_k = sum(x_{c,k})/N_c} \quad (2)$$

Table 1. List of keywords by topics

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
Law,	Network	Agriculture	Device	Science	Data	Culture	Secure
Public	System	Environment	Risk	Human	Learn	Art	Compute
Polite	Application	Treatment	Quality	Social	Mining	History	Mobile
Legal	Electron	Water	Regulatory	Life	Analysis	Philosophy	Data
Policy	Sensor	Microbiology	Compliance	Sociology	Machine	Literature	Cloud
Topic 9	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16
Educate	Cancer	Social	Biology	Image	Manage	Drug	Medicine
Learn	Oncology	Media	Biomed	Multimedia	Business	Pharmaceutical	Clinic
Teach	Pediatric	Behavior	Cell	Process	Economic	Pharma	Medic
Train	Vaccine	Community	Biotechnology	Interact	Market	Pharmacology	Emerge
Teacher	Dental	Issue	Molecular	Visual	Strategy	Pharmacy	Therapy
Topic 17	Topic 18	Topic 19	Topic 20	Topic 21	Topic 22	Topic 23	Topic 24
Health	Language	Chemic	Approach	Software	Engineering	Compute	Industrial
Care	Psychology	Chemistry	Platform	System	Material	System	Field
Healthcare	Linguist	Biochemical	Immunology	Model	Mechanic	Intelligent	Science
Primary	Region	Priority	Immune	Compute	Technology	Application	Expert
Pain	Anthropology	Toxicology	Virology	Application	Energy	Artificial	Innovate
Topic 25	Topic 26	Topic 27	Topic 28	Topic 29	Topic 30	Topic 31	
Disease	Engineering	Innovate	Science	Nurse	Application	System	
Diabetes	Technology	Practice	Nature	Nutrition	Physic	Digit	
Surgery	Civil	Advance	Environ	Adore	Mathematic	Service	
Cardiovascular	Electronic	Discuss	Animal	Sport	Science	Library	
Heart	Military	Trend	Veterinary	Beauty	Statist	Technology	

where $sum(x_{c,k})$ is the sum of x_k of all the conferences related to the k-th topic among all the conference held by country c and N_c is the total number of conferences in country c.

Based on the topic vector of each country, the concentration of country c can be determined by the standard deviation of the country's topic vector distribution, which is the degree of variance of the topic vector. The larger the standard deviation, the higher the degree of concentration for a particular research topic in the country.

Additionally, the calculated topic vectors for countries are also used to build a research similarity network. The research similarity network is expressed based on the similarities between countries' research topics. This provides important meanings because it shows a network of common interests and sharing and collaborations of research across countries. In addition, forming a cluster of countries of similar research topics can identify common interests of the countries and it can help forming a collaboration network between the clusters.

First, we used the cosine similarity for measuring the research similarities. The cosine similarity can be used to determine whether two vectors have the same direction, rather than measuring quantitative associations between vectors. In other words, it enables to identify if any two countries pursue similar research topics. The value of similarity between the two vectors will reside between 0 and 1, and higher the value is, the more similar two vectors are. The similarity of research topics between countries c and c' is defined as

$$\boldsymbol{Sim(V_c, V_{c'}) = \frac{\sum_{k=1}^{k=K} v_k v'_k}{\sqrt{\sum_{k=1}^{k=K}(v_k)^2}\sqrt{\sum_{k=1}^{k=K}(v'_k)^2}}} \quad (3)$$

Based on the results of calculating similarities between all countries, a research similarity network is constructed, where nodes represent countries and the edge thickness represents the degree of similarity. Then, we perform a modularity analysis on the constructed network. Modularity analysis is a method for detecting high-connectivity community structures among nodes and is well suited for cluster formation of countries based on research similarity [15]. In other words, it is possible to know how well each network is connected through optimized module analysis, and one network can be divided into a cluster called a module. The ratio of the links inside a group to the total number of links connected to an arbitrary group is referred to as an assortativity significance,

and a value that defines groups and determines how well the groups are divided is called a modularity [16].

A network can be represented as a matrix A_{ij} where the number of all connected links is m. The value of $\delta\left(t(node_i), t(node_j)\right)$ is defined as 1 if the characteristics of two nodes of $node_i$ and $node_j$ are the same, and it becomes 0 otherwise. The mean of assortative significance, T, is defined as

$$\boldsymbol{T} = \frac{\mathbf{1}}{\boldsymbol{m}} \sum_{\boldsymbol{i,j}} \frac{\boldsymbol{deg_i deg_j}}{\boldsymbol{m}} \boldsymbol{\delta}\left(\boldsymbol{t(node_i), t(node_j)}\right) \quad (4)$$

where deg_i means the degree of $node_i$.

Based on the assortative significance calculated by using Equation 4, the modularity, Q, is defined as

$$\mathbf{Q} = \frac{\mathbf{1}}{\boldsymbol{m}} \sum_{\boldsymbol{i,j}} \boldsymbol{A_{ij}}\, \boldsymbol{\delta}\left(\boldsymbol{t(node_i), t(node_j)}\right) - \boldsymbol{T} \quad (5)$$

The modularity is regarded as the difference between the mean assortative significance and the actually calculated value. Through time series analysis, we detect the changes in the international major research topics over time and identify the countries' leading research topic. Table 2 summarizes the list of major research topics based on the keywords in a selected topic list. We defined these with the help of the knowledge of an expert group.

Table 2. Topic groups and related topics

Names of topic group	Related topics
Law and Politics	Topic 1
Engineering and Technology	Topic 2, Topic 4, Topic 6, Topic 8, Topic 13, Topic 21, Topic 22, Topic 23, Topic 24, Topic 26, Topic 27, Topic 31
Physical and Life Sciences	Topic 3, Topic 12, Topic 19, Topic 30
Social Sciences and Humanities	Topic 5, Topic 7, Topic 11, Topic 18
Education	Topic 9
Health and Medicine	Topic 10, Topic 15, Topic 16, Topic 17
Business and Economics	Topic 14
Animal Sciences	Topic 28

IV. Experimental results

A. Analysis of research concentration by country

For the experiments, we collected information on the conferences, from 2015 to 2017, from an international academic conference search engine, called conference.city (http://www.conference.city). As for the number of conferences held by countries, the United States of America and the United Kingdom were ranked first and second, respectively, with more than 1,900 conferences from 2015. On the other hand, the number of conferences in Monaco and Ethiopia was retrieved below 9, which ranked about 100th. Therefore, the countries with less than 50 academic societies were considered to have no significant impacts on research trend analysis, and only 62 countries were selected for the experiments. Table 3 shows the list of selected countries by continents. Figure 1 depicts the current status of the number of conferences held in selected countries; darker region represents a larger number of conferences held in that area.

Table 3. List of selected countries according to continents

Continents	Countries
Africa	Egypt, Mauritius, Morocco, South Africa
Asia	Bahrain, Bangladesh, Cambodia, China, Hong Kong, India, Indonesia, Iran, Israel, Japan, Malaysia, Oman, Pakistan, Philippines, Qatar, Russia, Saudi Arabia, Singapore, South Korea, Sri Lanka, Taiwan, Thailand, Turkey, United Arab Emirates, Vietnam
Europe	Austria, Azerbaijan, Belgium, Bulgaria, Croatia, Czech republic, Denmark, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Luxembourg, Netherlands, Norway, Poland, Portugal, Romania, Slovakia, Spain, Sweden, Switzerland, Ukraine, United Kingdom
North America	Canada, Mexico, United States of America
Oceania	Australia, New Zealand
South America	Argentina, Brazil

Figures 1 and 2 show the distribution of research concentration by country and the number of conferences held. This can be explained by dividing into three groups in terms of research interests and diversity. The first group is defined as a leading group that has high interests in research because of the large number of conferences held and also accepts diversity of research with a low standard deviation. The second group is defined as a growing group in which the interests and diversity of research are continuously expanded. Finally, the third group is defined as a biased group because their research topics are focused on one side despite the low interests of that research.

The countries in the first leading group are represented in cross including the United States of America, United Kingdom, China, Germany, India, Spain, and France. These countries show quite active activities since they have held more than 500 conferences as shown in Figures 3 and 4. In addition, as their concentration scores are lower than 0.02, it can be said that they deal with diverse and various research topics rather than focusing only on few of particular topics. Especially, France has the lowest concentration score of

0.0086, which means that the most research fields are very actively covered.

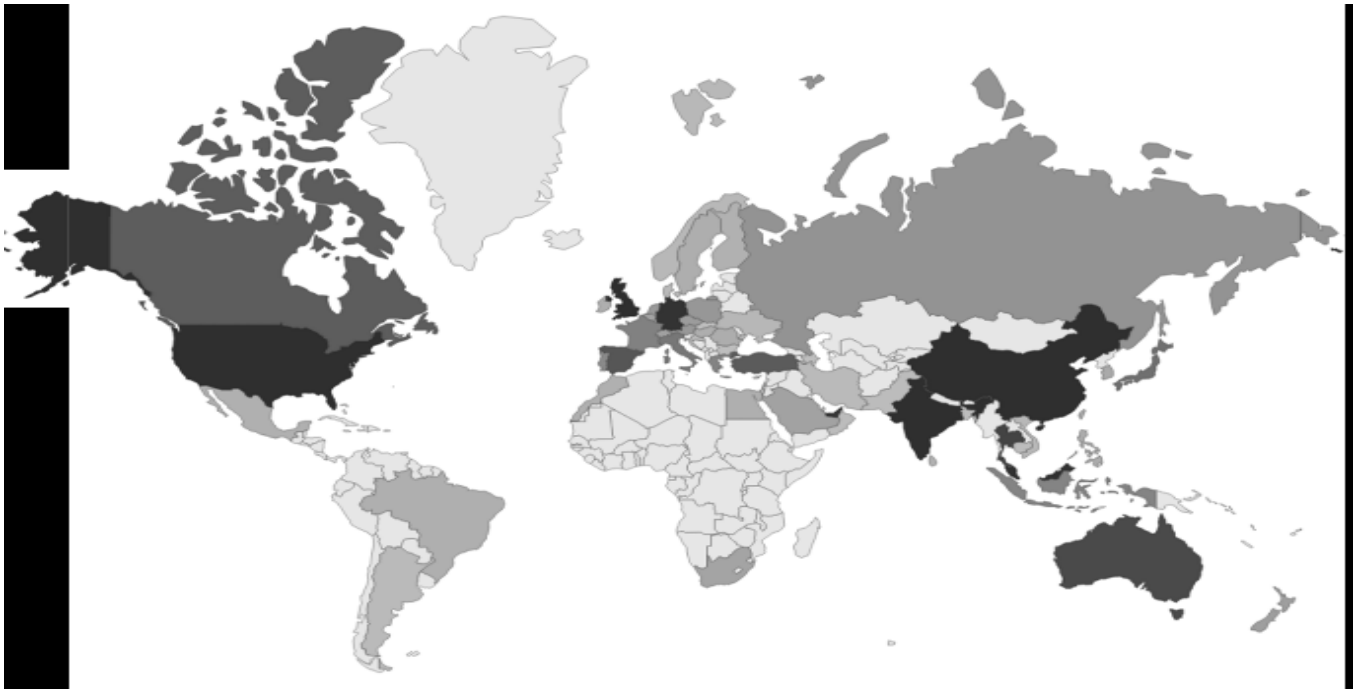


Figure 1. Visualization of the numbers of conferences collected according to countries. (The darker countries mean the more conferences held)

In addition, the United Kingdom has high interests in the field of pharmacy. In fact, 50 of the 100 most popular medicines in the world have been invented by the research institutes in the UK and about 400 pharmaceutical companies are in the country. This means that the pharmacy research has been significantly active in the UK. Meanwhile, India has particularly high level of research interests in statistics and engineering compared to the other countries.

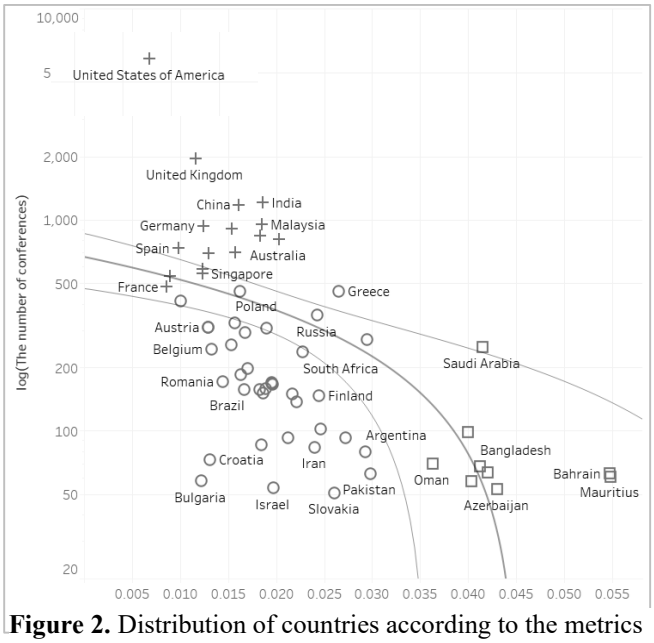


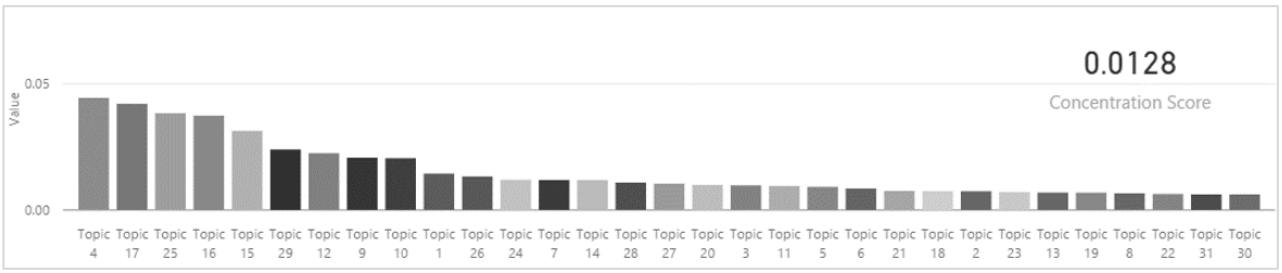
Figure 2. Distribution of countries according to the metrics considered.

Germany has a high concentration in the field of biotechnology. Germany is the second largest pharmaceutical biotech market in the world and has the largest market in Europe. According to the German Federal Ministry of Education and Research, about 48% of all biotechnology companies are focusing on the bio-healthcare industry and are focusing on new diagnostic technologies and new drug development.

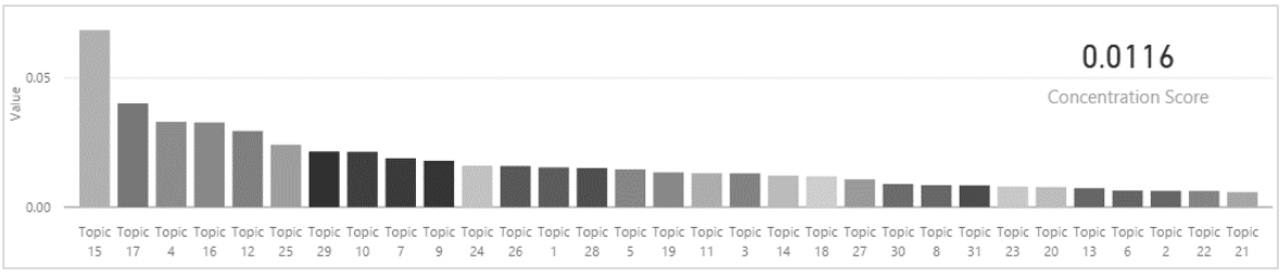
The countries in the growing group are represented in blue: Portugal, Poland, Switzerland, Austria, Indonesia, Belgium, Hong Kong, and etc. The distribution of the topic vector of those countries is shown in Figures 5, 6, and 7. Because the average number of conferences is more than 160 and the average concentration score is 0.02, the research interest and diversity are moderate.Especially Switzerland, Austria, and Hungary are highly interested in the medical field. Switzerland is more interested in healthcare, Austria pharmacy field, and Hungary therapeutics, respectively. In the case of Hungary, the pharmaceutical industry is one of the most outstanding and technologically advanced sectors in its economy, and it is the most developed in Central and Eastern Europe. Hungary is becoming a production base of bioscience enterprises expanding to Central and Eastern Europe, the Balkans, Eastern Europe, and Asia.

In addition, Asian countries such as Indonesia and Hong Kong are highly interested in the engineering field, and Belgium is more interested in the field of law.

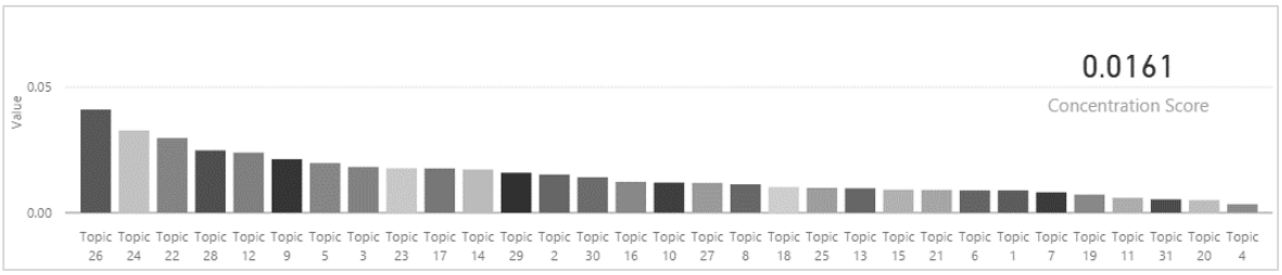
In particular, South Korea is also part of the growing group and based on the analysis of the topic vector. As shown in Figure 8, research is active in that the number of conferences held is more than 200, but the top 5 topics are concentrated on the engineering field, which indicates that diversity of research topics is needed. It is found that South Korea mainly focuses on the engineering field as shown in Figure 9. The top five topics are Topic 26, Topic 24, Topic 22, Topic 12, Topic 23, which are highly concentrated in the fields of chemistry, electronics, and industrial engineering, followed by biology and computer engineering.



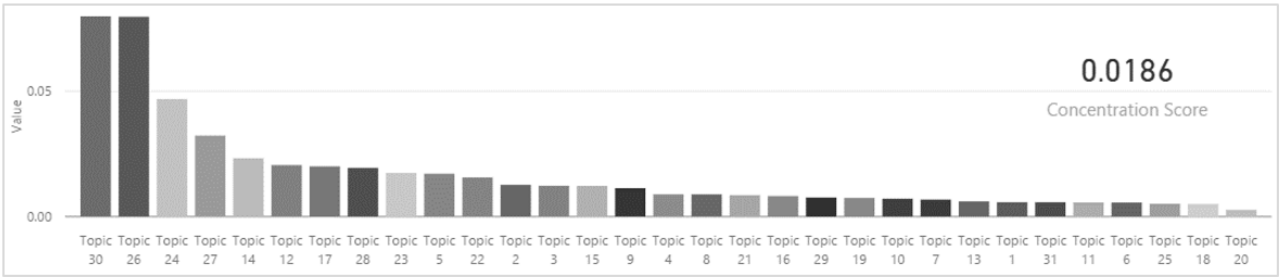
(a) United states of America



(b) United kingdom

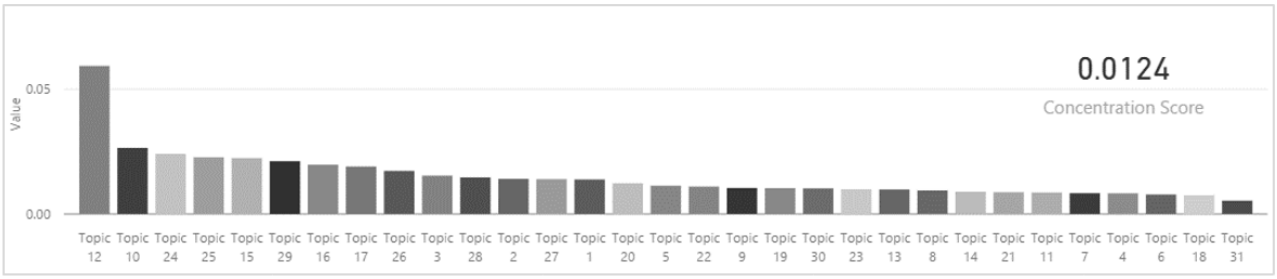


(c) China

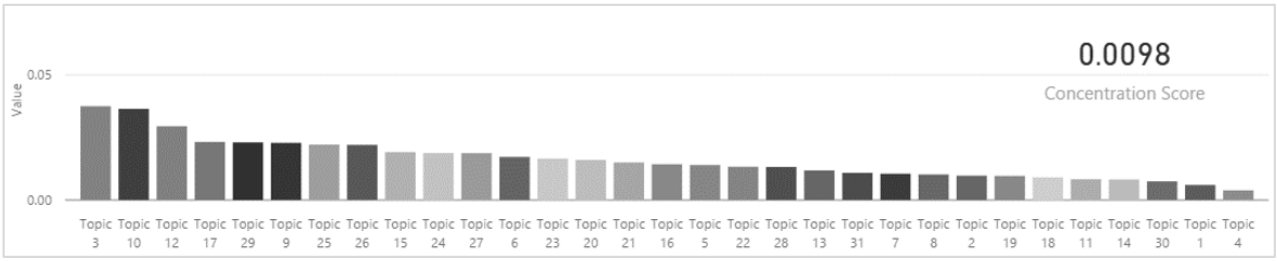


(d) India

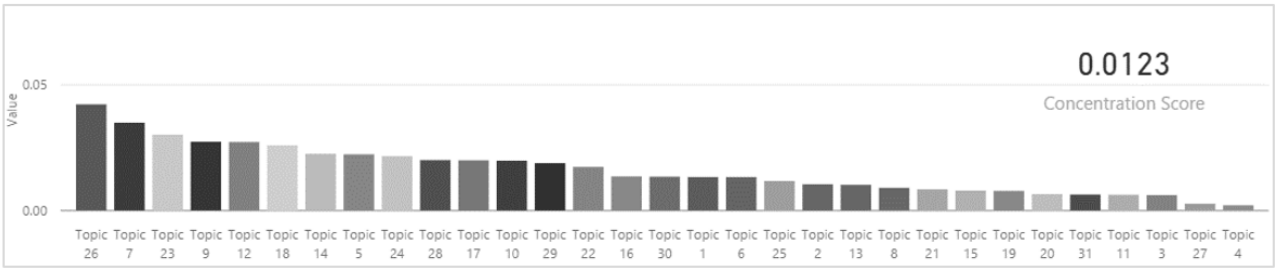
Figure 3. Distribution of topic vectors according to countries in leading group (1).



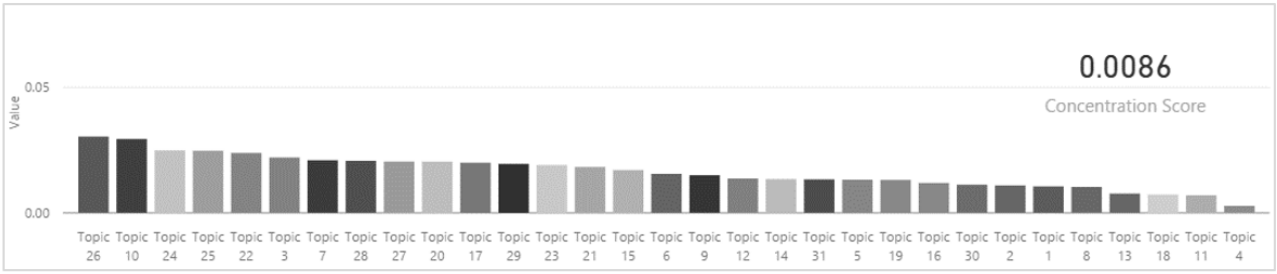
(a) Germany



(b) Spain

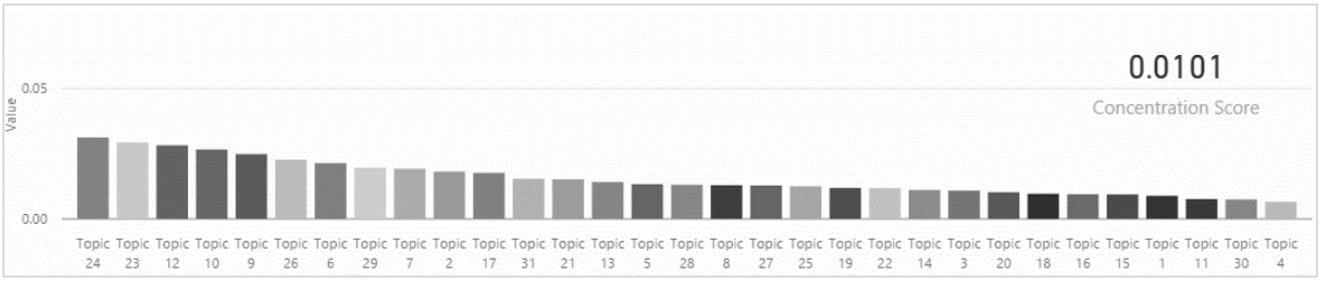


(c) Japan

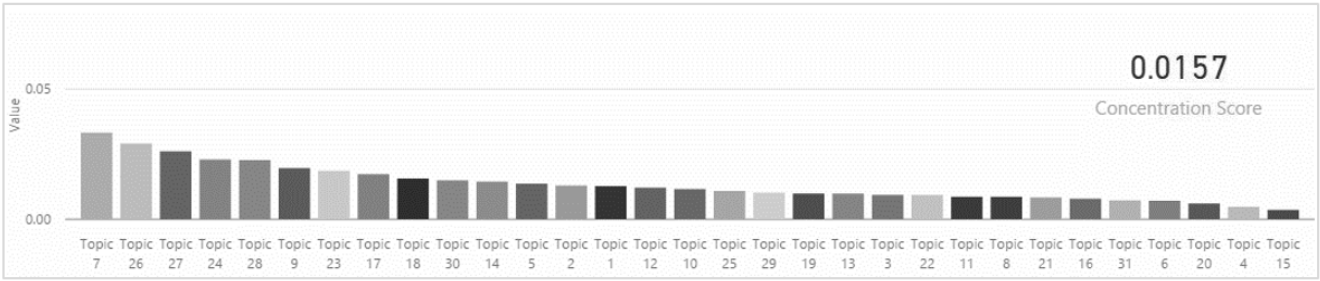


(d) France

Figure 4. Distribution of topic vectors according to countries in leading group (2).

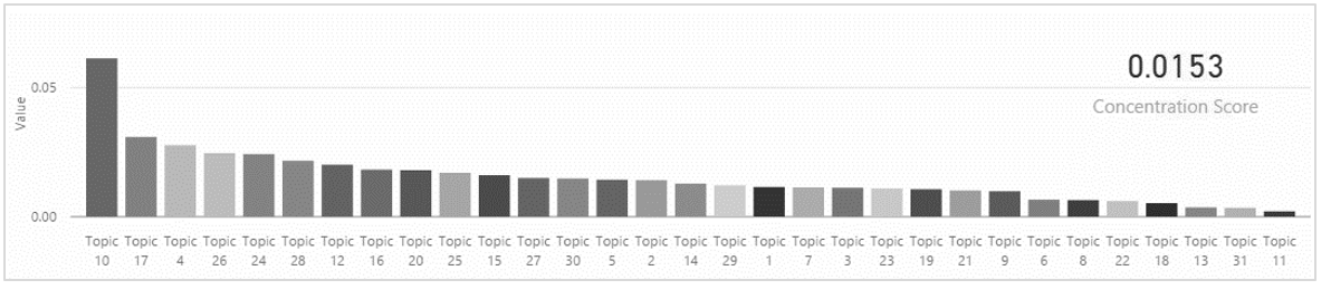


(a) Portugal

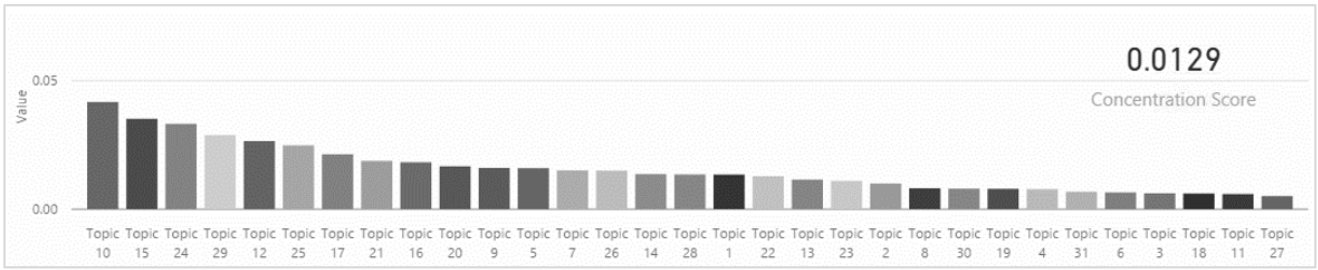


(b) Poland

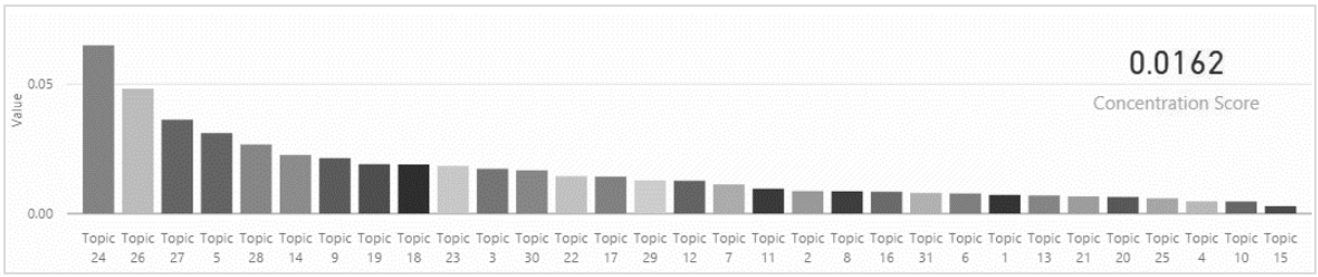
Figure 5. Distribution of topic vectors according to countries in growing group (1).



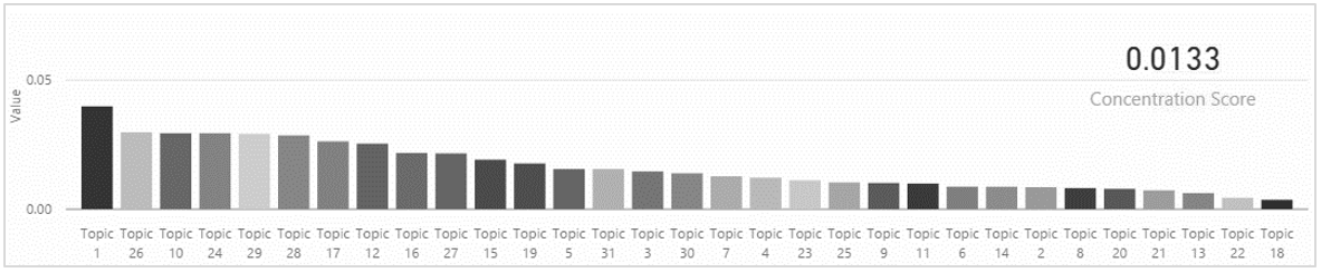
(a) Switzerland



(b) Austria

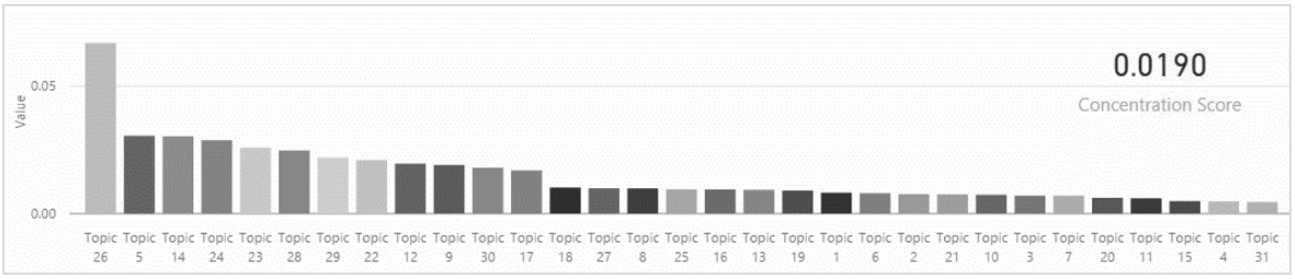


(c) Indonesia

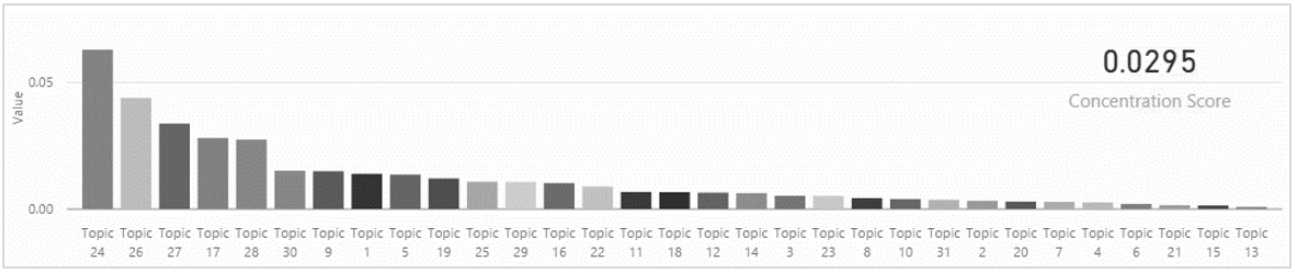


(e) Belgium

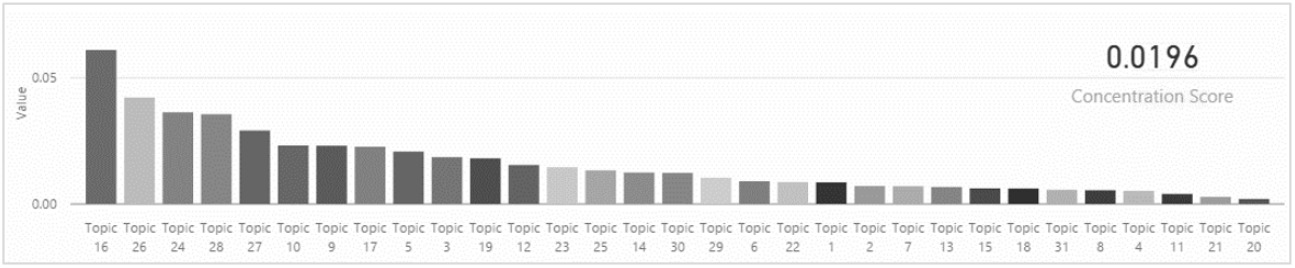
Figure 6. Distribution of topic vectors according to countries in growing group (2).



(a) Hong Kong



(b) New Zealand



(c) Hungary

Figure 7. Distribution of topic vectors according to countries in growing group (3)

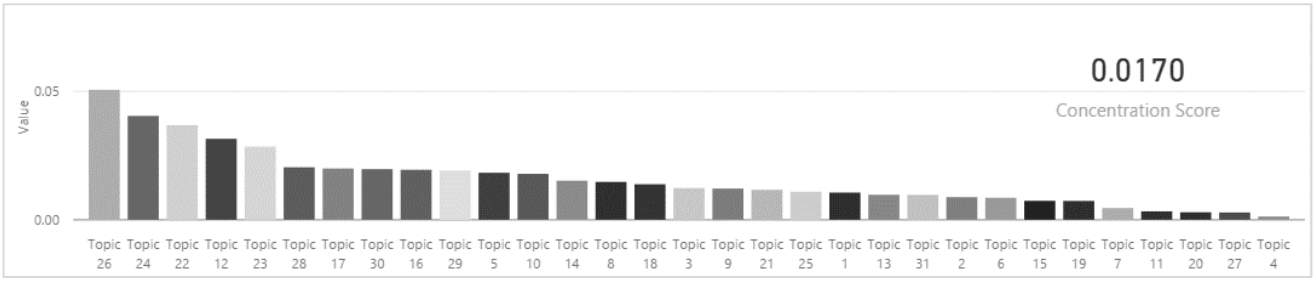


Figure 8. Distribution of South Korea topic vectors.

Countries in the third group, the biased group, are depicted as yellow: countries such as Mauritius, Azerbaijan, Cambodia, Bangladesh, and Saudi Arabia. Those countries are mainly in Southeast Asia and the continent of Africa. As shown in Figure 10, the top two topics are concentrated in engineering and technology fields: Topic 24 and Topic 26. However, it is difficult to say that the biased group has expertise in engineering because the engineering field is a general research field of 27% of all countries and the average number of conferences is just around 50.

Exceptionally, Cambodia has focused on natural science rather than engineering. The reason is that due to the conditions of developing country, Cambodia focuses on the urgent issues of the environment and agriculture than research on the engineering field, where research equipment and facilities are required. However, for more diverse research fields, Cambodia is making efforts to cooperate with various development cooperation partners, such as holding an educational research forum recently in cooperation with UNESCO. Furthermore, other countries in the biased group will also need to actively do research and expand the field through cooperative activities such as holding convergence forums or holding various conferences with countries of leading and growing groups.

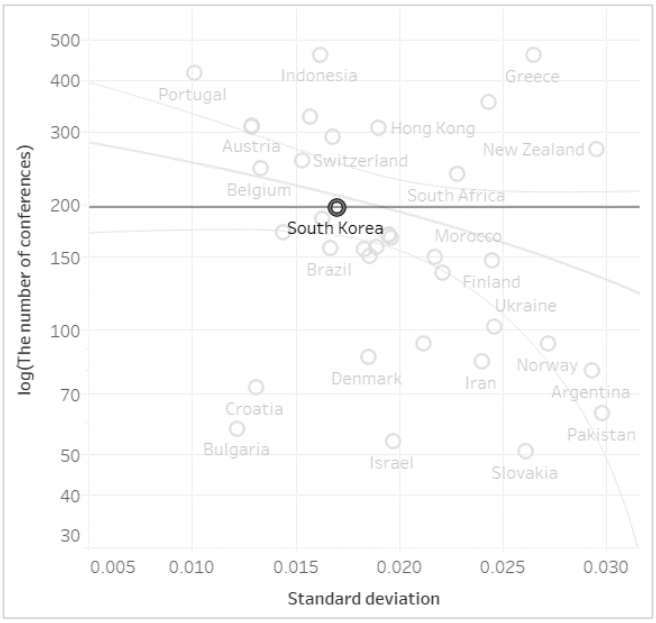


Figure 9. Position of South Korea compared to the other countries**.**

B. Analysis of relationships among countries

From the result of calculating the cosine similarity between the topic vectors of all countries, Table 4 shows the similarity results between countries with similarities greater than or equal to 0.927. Based on this, Figure 11 represents the research similarity network of all countries. Each node represents a country, and the darkness and edge thickness of the node means the weight of the research similarity between two connected nodes. The size of the node means the number of links considering the weight. If the weight of the edge is 0.80 or less, it is determined that the meaning of the similarity between the two nodes is insignificant and we removed the edge from the network. Finally, the number of nodes is 62 and the number of edges is 1,384. We decided the number of clusters as four because four-clusters has the highest modularity value of 0.044. Figure 12 shows the network of the clusters in which each node is colored according to its cluster.

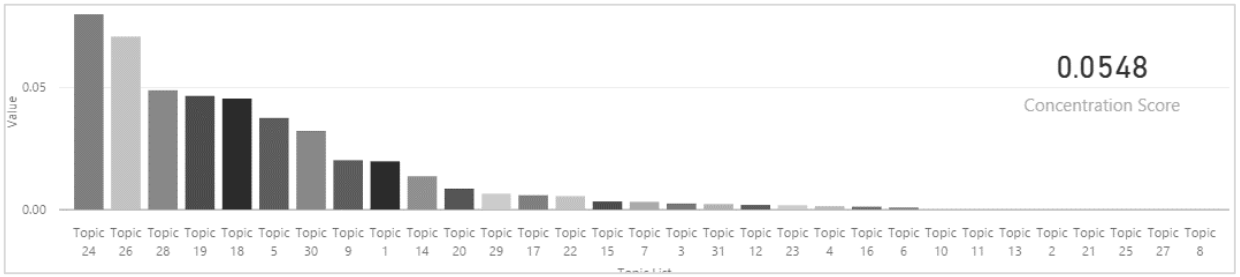
To analyze the research topic of the cluster, the top two topics are analyzed from the average of the topic vectors from each cluster as shown in Figure 13. In Figure 13, each cluster is represented by a class. The top two topics in class 0 are Topic 16 and Topic 28, and they include the words Medicine, Clinic, Medical, Energy, Therapy, Science, Nature, Environment, Animal, and Veterinary. This means that the class 0 is interested in research topics in medicine and natural sciences. In particular, class 0 has only two countries, Sri Lanka and Hungary, which are a relatively small amount compared to other classes. This is because Hungary, as described in Section 4.A, has excellent technology and research in the pharmaceutical industries and Sri Lanka pays more attention to the medical field because the prevalence and mortality of non-communicable diseases tend to increase.

Class 1 has many Southeast Asian countries such as the Philippines, Taiwan, Indonesia, Thailand, Iran, India, Cambodia, and Vietnam. In addition, the United Arab Emirates, Russia, South Africa, New Zealand, and Norway are belonging to class 1. The Topic 24 and Topic 26 are the top two topics in this class and the topics have the words Industry, Field, Science, Expert, Innovation, Engineering, Technology, Civil, Electronic, and Military. So, class 1 can be said to represent the industry and engineering fields.

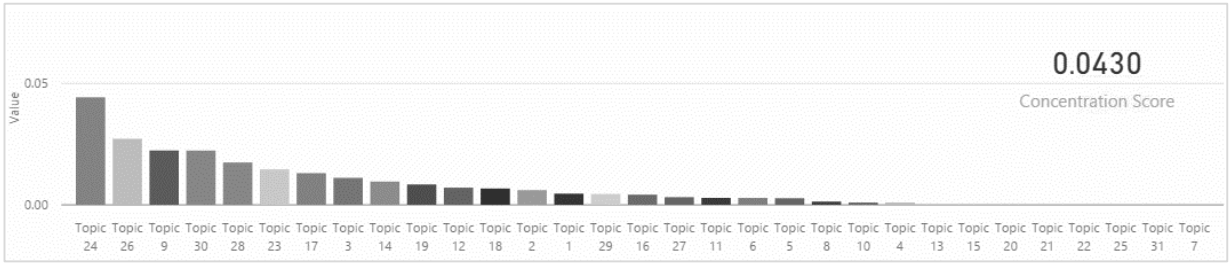
Class 2 includes major Asian countries such as China, Japan, Hong Kong, South Korea and Singapore, and Eastern Europe and neighboring countries such as Poland, Slovakia, Romania, Greece, and Turkey. The top two topics are Topic 9 and Topic 26, which has the words Industry, Field, Science, Expert, Innovation, Engineering, Technology, Civil, Electronic, and Military. Therefore, class 2 also can be said to represent the industry and engineering fields.

Finally, most of the European countries such as Austria, Czech Republic, Netherlands, United Kingdom, Spain, Italy, Germany, Portugal, France, Sweden, and Switzerland, and North and South American countries like Mexico, Brazil, Canada, and the USA are belonging to class 3. The top two topics are Topic 10 and Topic 12, which have the word of Cancer, Oncology, Pediatric, Vaccine, Dental, Biology, Biomedical, Cell, Biotechnology, and Molecular. Therefore, we can determine that countries in Europe and the Americas actively do research in those fields compared to other countries.

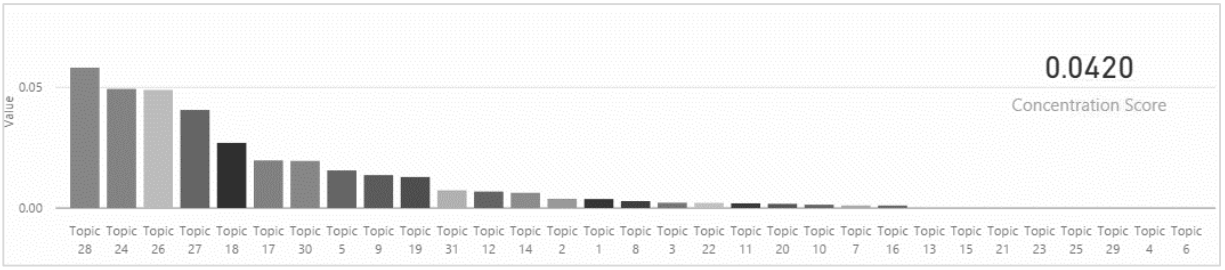
Since each cluster represents different research topics, it would be possible to strengthen the collaborations between research communities with different subjects by holding convergence forums or conferences for research collaboration among countries. Therefore, through this study on conference information, we can contribute to providing diverse research opportunities for different community and countries. The similar research groups around the world can be seen at a glance from Table 5 and Figure 14; Table 5 shows the list of the top two topic keywords for each class and Figure 14 shows the classes with different darkness on a map. As you can see in Figure 14, the countries in each cluster are located in similar geographical locations. This suggests that the geographical factor affects the similarity of research.



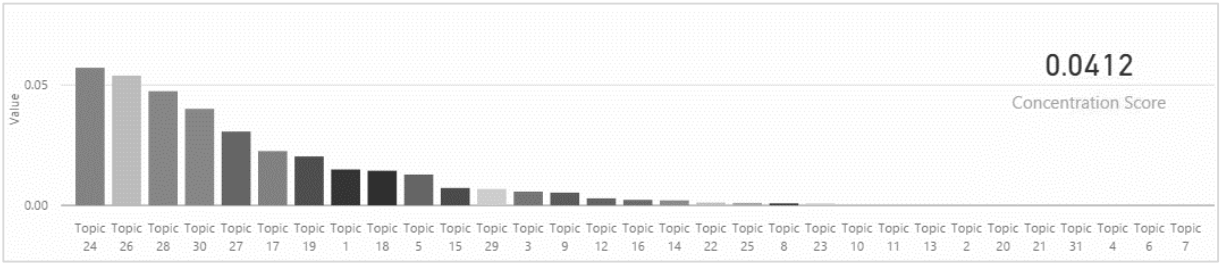
(a) Mauritius



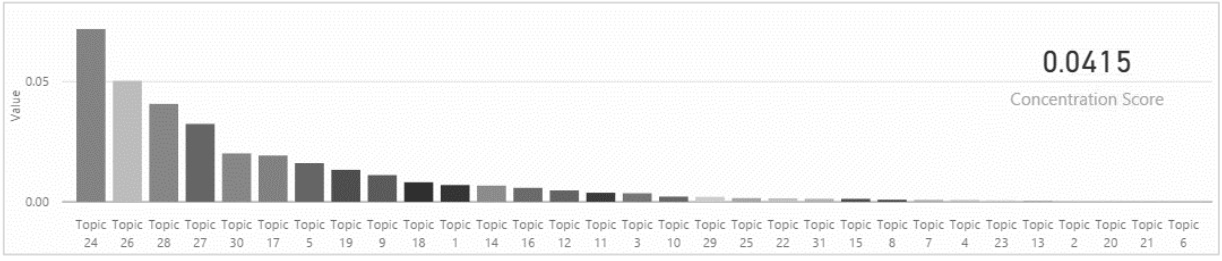
(b) Azerbaijan



(c) Cambodia



(d) Bangladesh



(e) Saudi Arabia

Figure 10. Distribution of topic vectors according to countries in bias group.

Table 4. Cosine Similarities between Countries (≥ 0.927).

Source	Target	Cosine similarity
Bahrain	Saudi Arabia	0.976374
Saudi Arabia	Bahrain	0.976374
Poland	Romania	0.967933
Romania	Poland	0.967933
China	South Korea	0.967036
South Korea	China	0.967036
New Zealand	Saudi Arabia	0.967005
Saudi Arabia	New Zealand	0.967005
Indonesia	Taiwan	0.961247
Taiwan	Indonesia	0.961247
Bangladesh	Saudi Arabia	0.959624
Saudi Arabia	Bangladesh	0.959624
Bahrain	New Zealand	0.958869
New Zealand	Bahrain	0.958869
Hong Kong	Philippines	0.945785
Philippines	Hong Kong	0.945785
Bangladesh	Cambodia	0.944614
Cambodia	Bangladesh	0.944614
Cambodia	Saudi Arabia	0.943905
Saudi Arabia	Cambodia	0.943905
Austria	Netherlands	0.943695
Netherlands	Austria	0.943695
Greece	Portugal	0.942445
Portugal	Greece	0.942445
Greece	Japan	0.942105
Japan	Greece	0.942105
France	Spain	0.940514
Spain	France	0.940514
China	Hong Kong	0.940239
Hong Kong	China	0.940239
Australia	Canada	0.939758
Canada	Australia	0.939758
Australia	Belgium	0.939457
Belgium	Australia	0.939457
Luxembourg	Saudi Arabia	0.938512
Saudi Arabia	Luxembourg	0.938512
Bahrain	Luxembourg	0.937042
Luxembourg	Bahrain	0.937042

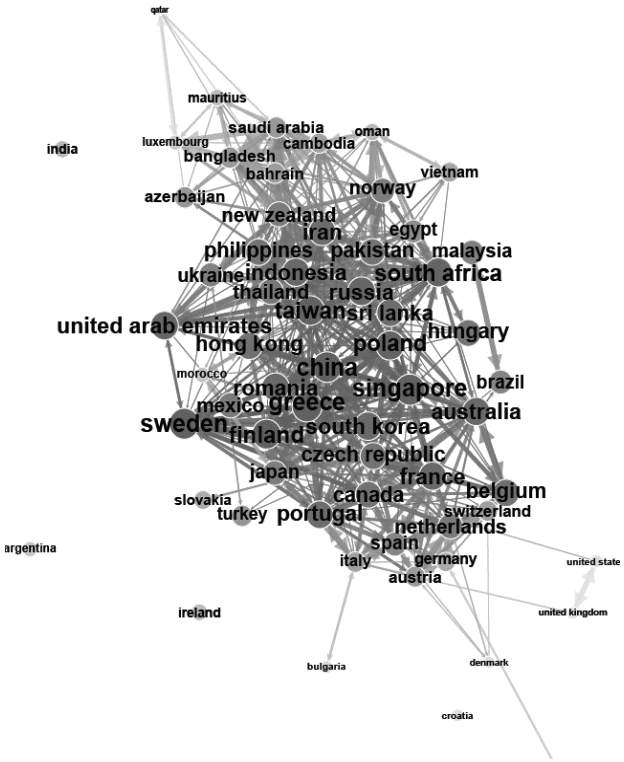


Figure 11. Network based on cosine similarity.

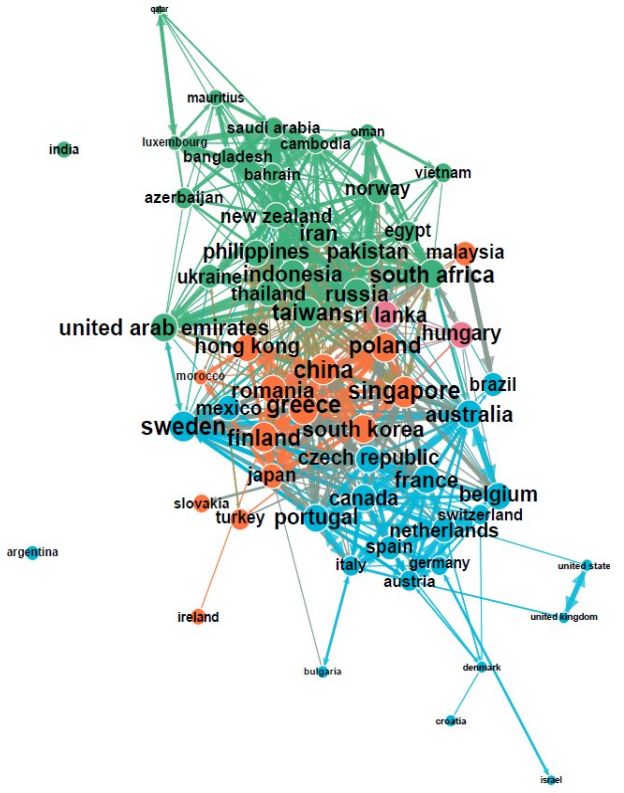


Figure 12. Network results with modularity analysis. (The colors are different depending on the modularity class.)

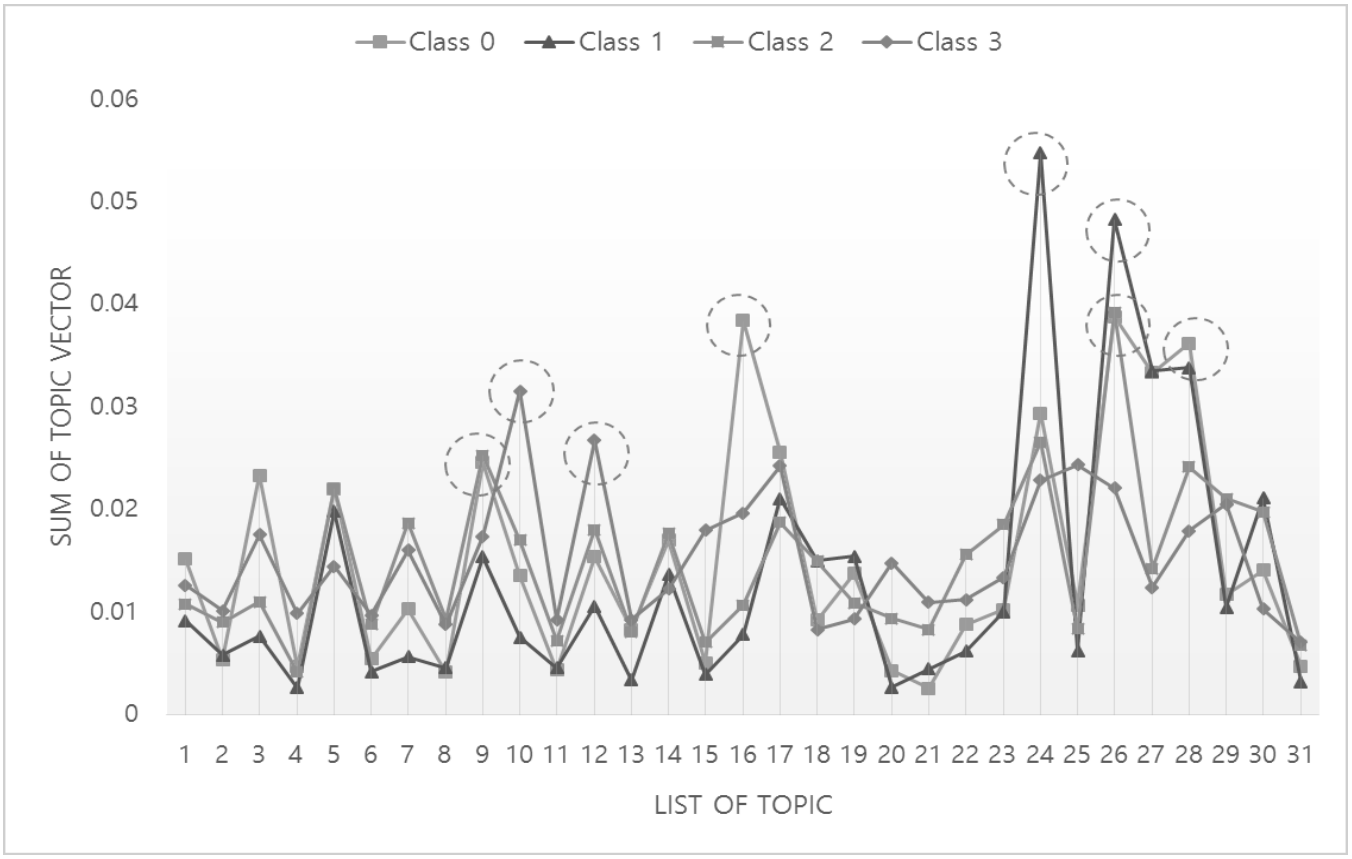


Figure 13. Distribution of the average of topic vectors according to each class. (The dotted circle is the top 2 of all topics.)

Figure 14. Map of country distribution according to modularity class.

Table 5. The topics in each class and their keyword lists.

Class	Representative topics	List of keywords
0	Topic 16	Medicine, Clinic, Medical, Energy, Therapy
	Topic 28	Science, Nature, Environment, Animal, Veterinary
1	Topic 24	Industry, Field, Science, Expert, Innovation
	Topic 26	Engineering, Technology, Civil, Electronic, Military
2	Topic 9	Educate, Learn, Teach, Train, Teacher
	Topic 26	Engineering, Technology, Civil, Electronic, Military
3	Topic 10	Cancer, Oncology, Pediatric, Vaccine, Dental
	Topic 12	Biology, Biomedical, Cell, Biotechnology, Molecular

V. Conclusions

In this research, we suggest analytical methods for understanding the research trends and the relationships among countries by utilizing topic mining and network analysis techniques.

For the international research trend analysis, we proved that meaningful results can be derived from the information on 25,000 conferences held from 2015 to 2017. Through topic modeling, a total of 31 topic vectors were generated to identify the concentration of research by country. In addition, the network of countries with research similarity was constructed and analyzed. Then, research trends by year and country were investigated by analyzing the change of topic distribution. The results are summarized as follows.

First, based on the analysis of the research concentration, each country was classified into either leading group, growth group, or bias group according to the degree of research interest and diversity. The leading groups were mainly developed countries such as the United States of America, United Kingdom, China, and more than 500 conferences were held per year on average. Also, those countries have very low concentration scores and this showed that they have diverse research activities. In the growing group, there are countries such as Portugal, Poland, and Switzerland, and the interest and diversity of research were reasonable. Especially Switzerland, Austria and Hungary were particularly interested in medicine and Belgium was highly interested in law. Finally, most of the countries in the biased group are from Southeast Asia and Africa and we can find that they concentrate in the engineering field.

Second, as a result of clustering of countries based on the research similarity network, four clusters were created, and each cluster confirmed the existence of meaningful connections among research topics. The first cluster focused on medical and natural sciences, the second on industrial and engineering, the third on education and engineering, and the fourth on biomedical sciences, respectively.

Finally, the results of the time series analysis showed that the interest in engineering and technology was overwhelming, and the fields of health and medicine, law and politics, and education had an overall downward trend. Also, India has shown relatively high interest in the fields of physics and life sciences, business, and economics.

Acknowledgment

This work was supported by Incheon National University Research Grant in 2016.

References

[1] J.A.S. Almeida, A.A.C.C. Pais, and S.J. Formosinho. "Science Indicators and Science Patterns in Europe", *Journal of Informetrics*, 3(2), pp. 134–142, 2009.

[2] J.H. Jin, S.C. Park, and C.U. Pyon. "Finding Research Trend of Convergence Technology based on Korean R&D Network", *Expert Systems with Applications*, 38(12), pp. 15159–15171, 2011.

[3] K. Oh, and M. Lee. "Research Trend Analysis of Geospatial Information in South Korea Using Text-Mining Technology", *Journal of Sensors*, 2017, pp. 1-16, 2017.

[4] X. Xin, J. Li, J. Tang, and Q. Luo. "Academic Conference Homepage Understanding using Constrained Hierarchical Conditional Random Fields". In *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM)*, pp. 1301-1310, 2008.

[5] I. Muhammad, H. Fredrik, A. Muhammad, and Z.A. Jahan. "Recent Research Trends in Organic Rankine Cycle Technology: A Bibliometric Approach." *Renewable and Sustainable Energy Reviews*, 81(1), pp. 552-562, 2018.

[6] Z. Zhuang, E. Elmacioglu, D. Lee, and C.L. Giles. "Measuring Conference Quality by Mining Program Committee Characteristics". In *Proceedings of the ACM IEEE Joint Conference on Digital Libraries*, 2007.

[7] Y.-H. Tseng, C.-J. Lin, Y.-I Lin, "Text mining techniques for patent analysis," *Information Processing & Management*, 43(5), pp. 1216-1247, 2007.

[8] J. Tang, B. Wang, Y. Yang, P. Hu, Y. Zhao, X. Yan, B. Gao, M. Huang, P. Xu, W. Li, A. K. Usadi, "PatentMiner: topic-driven patent analysis and mining," *In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '12)*, pp. 1366-1374, 2012.

[9] C. Pan, W. Li, "Research paper recommendation with topic analysis," *International Conference on Computer Design and Applications*, pp. V4-264-V4-268, 2010.

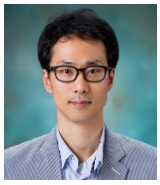
[10] T. L. Griffiths, M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, 101(Supple 1), pp. 5228-5235, 2004.

[11] K. Rajaraman, A.-H. Tan, "Topic Detection, Tracking, and Trend Analysis Using Self-Organizing Neural Networks," *Advances in Knowledge Discovery and Data Mining*, pp. 102-107, 2001.

[12] X. Wang, A. McCallum, "Topics over time: a non-Markov continuous-time model of topical trends," *In*

Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '06). pp. 424-433, 2006.
[13] D.M. Blei, A.Y. Ng, and M.I. Jordan. "Latent Dirichlet Allocation" *Journal of machine Learning research*, pp. 993-1022, 2003.
[14] T. Cho, and J.H. Lee. "Latent keyphrase extraction using LDA model", *Journal of Korean Institute of Intelligent Systems*, 25(2), pp. 180-185, 2015.
[15] M.E. Newman. "Modularity and community structure in networks". In *Proceedings of the National Academy of Sciences*, 103(23), pp. 8577-8582, 2006.
[16] R. Zafarani, M.A. Abbasi, and H. Liu. *Social Media Mining: An Introduction*, Cambridge University Press, United Kingdom, 2004.

Author Biographies



Kwanho Kim received his Ph.D. degree in industrial engineering from Seoul National University, Korea, in 2012. He is currently an associate professor of the Department of Industrial and Management Engineering in Incheon National University (INU), Korea. His research interests include statistical methodologies to analyze and manage information such as machine learning, text mining, information retrieval, recommendation systems for various business fields ranging from manufacturing to mobile business.

Sue-Kyung Lee received her B.S. and M.S. degrees in industrial management and system engineering from Incheon National University (INU), Korea, in 2016 and 2018, respectively. She is currently a data scientist in SK C&C in Korea. Her research areas are text mining, relation analysis, topic extraction based on machine learning and deep learning technologies.

Heemin Park received the B.S. and M.S. degrees in computer science from Sogang University, South Korea, in 1993 and 1995, respectively, and the Ph.D. degree in electrical engineering from the University of California, Los Angeles, in 2006. He was with Yonsei University, Sookmyung Women's University, and Samsung Electronics, South Korea. He is currently an Associate Professor with the Department of Software, Sangmyung University, Cheonan, South Korea. His research interests include intelligent systems, networked and embedded computing systems, cyber physical systems, multimedia applications, and entertainment computing.

Jinseok Chae received the B.S., M.S., and Ph.D. degrees in computer engineering from Seoul National University, South Korea, in 1990, 1992, and 1998, respectively. In 2006, he joined the Department of Computer Science, California State University San Bernardino, California, USA, as a Visiting Scholar. From 2012 to 2014, he was the Dean of Admissions and Student Affairs with Incheon National University, South Korea. From 2015 to 2017, he was a Visiting Scholar with the Department of Computer Science, Texas Tech University, TX, USA. He was with the Engineering Laboratory, Seoul National University, and also with the Korea Research Information Center, South Korea. He is currently a Professor with the Department of Computer Science and Engineering, Incheon National University. His research interests include Internet software, Web technology, and Mobile computing. He was an Editor-in-Chief of the Journal of KIISE: Computing Practices and Letters from 2011 to 2014.