International Journal of Engineering & Technology, 7 (2.33) (2018) 1222-1228

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Extracting Environmental Research Trends using LDA

Do-Yeon Kim[1] and Sung-Won Kang[2*]

[1,2]Korea Environment Institute, 370, Sicheong-daero, Sejong-si, Republic of Korea, 30147
**Corresponding author E-mail: swkang@kei.re.kr*

Abstract

In this study, we compared topics of two distinct text data on environmental issues: environmental research reports and on-line environmental news using a Latent Dirichlet Allocation (LDA) analysis. For the environmental research reports, we used digitized research reports from the Korea Environment Institute, whereas for the newspaper, we crawled environment news articles on the Naver portal service. Once we extracted the topics, we compared the annual share of each topic in each text medium. From the LDA analysis, ten topics emerged from each medium. Six are common in both media, whereas four of the latest issues, namely, "gene variation," "noise," "health," and "data," only appeared in environmental news. In addition, among the six common topics, the share of "water pollution" and "waste" topics in environmental news appears to lead the share of the same two topics in the environmental research reports. These two results suggest that research topics tend to fall behind environmental issues in terms of latest interest. This study raises the possibility of using the LDA model to analyze research trends and find new research topics.

Keywords: *Environmental Research Trends, Machine Learning, Natural Language Processing, LDA, Topic Model*

1. Introduction

While qualitative analyses such as a literature study or expert assessment have previously been used to select research topics, such methods are likely to involve too much time for the results obtained. In addition, existing content analysis is in danger of mirroring subjective values and personal views of researchers, and it is difficult to understand large amounts of data. This study aimed to remedy the weak points of a quantitative analysis and an existing content analysis by applying a topic clustering technique. There is no domestic research comparing the environmental research trends with other media using a topic clustering technique. Therefore, in this study we conducted an analysis using the Latent Dirichlet Allocation (LDA) model to complement the disadvantages of a quantitative analysis. For an analysis of the environmental research trends, we collected policy research reports by the Korea Environment Institute (KEI). In addition, for a comparison with other media with different characteristics, on-line news Naver news articles from 193,636 articles were collected using the Java HTML Parser, jsoup.

This study focused on an LDA analysis for each medium to examine the major topics. For the analysis, we first constructed an LDA model. The analysis findings were visualized using the LDAvis package provided by R. A trend analysis was conducted to compare and analyze the trends of the major topics of each medium. Finally, we investigated the trends of common topics and the individual topics of each medium. In this study, we examined the 24-year KEI research trends using the topic clustering technique, and analyzed whether the KEI research trends corresponded to the social needs for environmental research. In addition, this study aimed at providing reference data that are necessary for researchers to develop new research topics by forecasting future environmental research trends based on the analysis results.

2. Materials and Methods

2.1. *LDA Topic Modeling*

As a probabilistic graphical model suggested by Blei(2008), LDA is used to model the probability of including words on certain topics using a Dirichlet distribution [1].In other words, the LDA model uses a latent probability estimation technique, which is premised on the idea that a document may include several topics, or several documents may share common topics [2]. Each topic has a certain distribution, and the distribution of each topic has features of the hierarchical model consisting of subsets of the entire data distribution [1].

This LDA model is an unsupervised learning method, which does not require any prior information of the documents or control. For this reason, it has been spotlighted in the field of natural language processing and machine running. Most of the topic models used recently are based on LDA. LDA was developed as a text mining tool, but this is currently used as an analysis tool in different fields applying images, social networks, genetic information, and so on. An expression of the early LDA model is shown in Figure 1 [3].

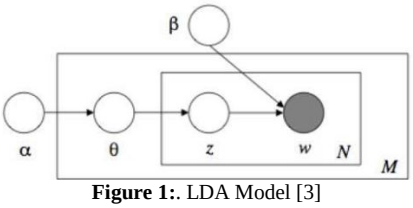


Figure 1:. LDA Model [3]

In Figure 1, θ presents the probability distribution of topics regarding a particular document, and z presents a probability distribution of words on topics as a conditional probability for θ. In addition,αindicates a Dirichlet prior probability of the topics in each document, and β is a constant before modeling as a Dirichlet

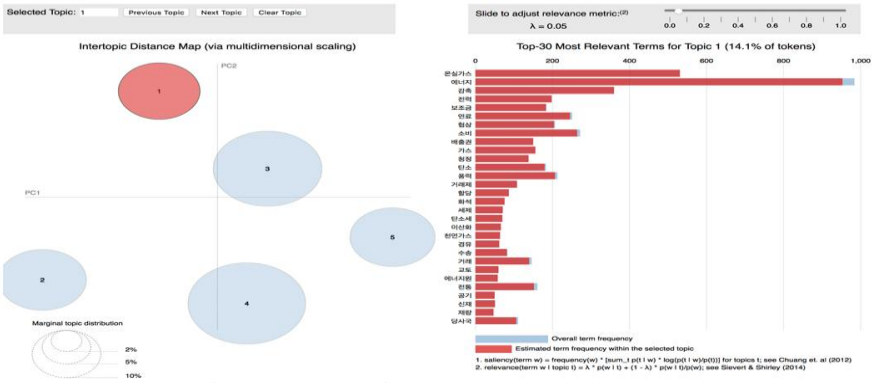
prior probability of word distribution per topic. Using the parameter values of α and β, which are defined in advance in the given document set, θ and z are probabilistically estimated, and a word 'w' is finally estimated by selecting a topic through a topic distribution in a document, and a word through a word distribution within a topic. In other words, LDA is a type of generative model, and when the LDA model is learned first, the number of topics and the prior probability distribution must be determined manually[4].

2.2. Research Using the LDA Technique

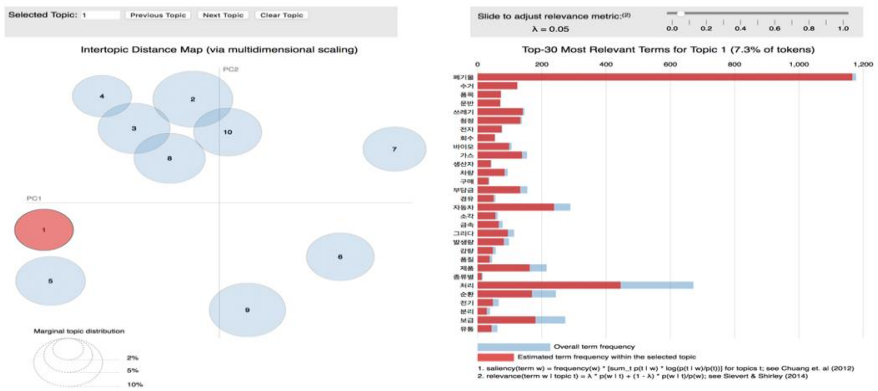
Using the LDA-based topic clustering technique to examine the topics and trends of specific fields in massive amounts of text data will make it possible to classify topics of documents objectively, and to analyze the trend of each topic. Because of these advantages, topic modeling has been used as an analysis tool in research of different fields using literature as the research data. Domestic and foreign studies using an LDA-based topic clustering technique to analyze the trends of specific topics are as follows. Park (2015) performed topic modeling after selecting the top-five highest rated dramas in 2014 and crawling documents written on a day when the five dramas were aired. He then analyzed the topic trends of the five dramas along with their ratings [5]. Wang and McCallum (2006) applied the Topics over Time(TOT) of the LDA topic models to examine how the topics changed over time. For the analysis, Neural Information Processing System(NIPS) conference proceedings from 1987 to 2003 and 200-year data of U.S. presidential speeches were employed. By applying the TOT model to each text, topics were discovered, and the trend of each topic was analyzed [6]. Newman and Block(2006) extracted topics from 18-century newspaper texts through topic modeling, and analyzed how each topic changed over time in order to comprehend early American society and its publication culture [7].Gerrish and Blei(2010) analyzed the content changes of topics over time in a corpus of papers using the dynamic topic model, and applied it to measure the influence of individual studies[8]. As stated above, it was revealed that topic modeling has been mostly used for a trend analysis in specific fields, and there have been no cases apply it to environmental topics. In this study, we extracted topics from research reports published by an environmental policy institute and from environmental news articles on Naver using the LDA model and analyzed the chronological changes of the issues and their aspects comprehensively.

3. Results and Discussion

In natural language processing, we extracted topics using LDA-based topic modeling functions and selected 30 keywords related to each topic. Based on the topic modeling results, we also implemented 2D visualization, as well as visualization of the major keyword probability distribution category using the LDAvis package provided by R. In relation to LDA visualization, the left graph expresses five topics using circular shapes, and abbreviates them into two dimensions (PC1, PC2). The width of each circle represents a ratio to all N tokens in a corpus. The right side presents a bar graph to show the most important 30 words in each topic. The red bar means the frequency of terms created by the given topics, and the blue bar refers to the overall frequency of each term in a corpus [9]. The specific gravity of the two bars is adjusted, based on the gamma value (λ)on the upper right. In this study, a small gamma value of 0.05 was set to extract words with a high specific gravity of the red bar. In other words, we extracted the keywords that frequently appeared only in specific topics, not common topics, and decided on the themes of the topics based on the features of the keywords extracted. Figure 2 below shows part of the LDAvis visualization analysis results based on the KEI research reports and Naver environmental news.



(a) The first topic extracted from the KEI research reports: Energy resources



(b) The first topic extracted from Naver environmental news: Waste

Figure 2:. Part of LDAvis results in each medium

3.1. *LDA Analysis of KEI Research Reports*

3.1.1. Topic Clustering and Keyword Analysis Results of Each Topic

According to the LDA analysis results of the KEI research reports, there were a total of six topics (energy resources, waste, external cooperation, water environment/environmental impact assessment, and climate change). Table 1 shows five topics and 16 major keywords of each topic. The fourth topic (water environment/environmental impact assessment) occupied the largest proportion (29.3%) of all topics, and the first topic (energy resources) occupied the smallest proportion (14.1%) of all topics.

Table 1: Five topics related to KEI research reports and relevant keywords

No.	1	2	3	4	5
Topic Title	Energy Resources	Waste	External Cooperation	Water Environment/ Environmental impact assessment	Climate Change
%	14.1	16.4	25.0	29.3	15.3
1	Greenhouse gas	Waste	Cooperation	Underground water	Climate
2	Energy	Noise	Forum	Environmental impact assessment	Change
3	Electric power	Chemistry	North Korea	Stream	Drought
4	Fuel	Processing	water and sewage	Wetland	Disaster
5	Gas	Waste water	The two Koreas	Purification	Heat wave
6	Clean	Sewage	Northeast Asia	Topography	Sea level
7	Carbon	Collection	Exchange	Waterside	Temperature
8	Wind power	Harmful	Continuous	Water source	Flood waters
9	Detergent	Block	Vietnam	a building site	Death
10	Carbon tax	Incineration	Citizen	A place of residence	Natural disaster
11	Carbon dioxide	Trash	South andNorth Korea	HangangRiver	Ozone
12	Natural gas	Bad smell	Initiative	Aquatic	Flood
13	Diesel	Burden	Governance	Soil	Typhoon
14	Air	Processing site	Philippines	Green	Precipitation
15	Renewable energy	Harmful	Asia	Saemangeum	Temperature

3.1.2. Analysis of KEI Research Report Trend of Each Topic

To analyze the research trends of the five topics extracted from the KEI research reports from 1993 to 2016 in detail, we classified the 14th to 18th presidential terms into four periods (1993 to 2002, 2003 to 2007, 2008 to 2012, 2013 to 2016). The 14th president Kim Young-Sam and the 15th president Kim Dae-Joong served from 1993 to 2002. During this period, KEI was founded, and as the initial stage of research and development, there was a small number of research reports published by KEI, which resulted in no big difference in the analysis results even when the period was subdivided. Therefore, the two presidential terms were combined into a single period. Figure 3 shows a graph on the KEI research report trends. The four periods in Figure 3 are differentiated by the shadow of the graph background.

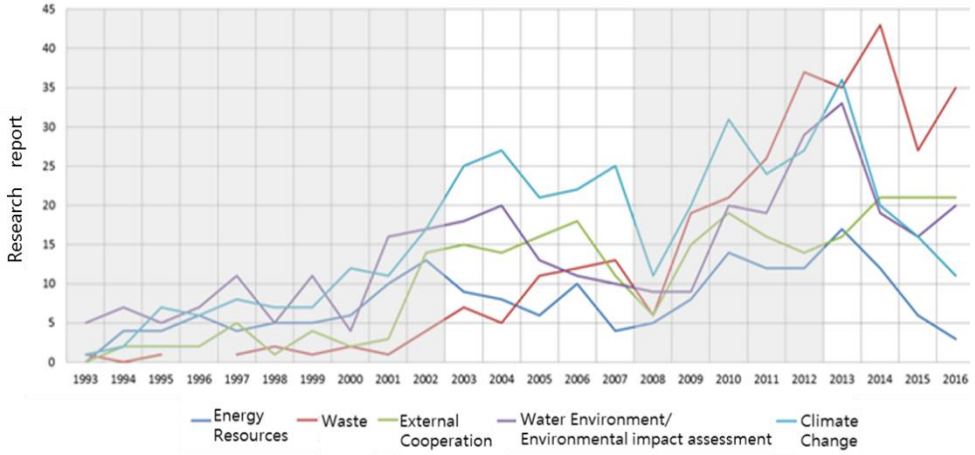


Figure 3:. KEI research trends of each topic

The research trends of each topic are similar overall from 1993 to 2002. From 2003 to 2007, there were active studies on climate change, and studies on water environment/environmental impact assessment and energy resources declined. From 2008 to 2012, studies on waste and water environment/environmental impact assessment rapidly increased. From 2013 to 2016, there were active studies on waste, and starting from 2015, the amount of research was reduced.

3.2. *LDA Analysis of Naver Environmental News*

3.2.1. Topic Clustering and Keyword Analysis Result of Each Topic

According to the LDA analysis results of Naver environmental news, there were a total of ten topics (waste, environmental impact assessment, energy resources, gene variation/noise, water pollution, sea/wind power, health/data, international cooperation, water environment, and climate change). Table 2 indicates 15 major keywords of the ten topics. The second topic (environmental impact assessment) occupied the largest proportion (13.4%) of all topics, and the first topic (waste) occupied the smallest proportion (7.3%) of all topics.

Table 2: Ten topics related to Naver environmental news and relevant keywords

No.	1	2	3	4	5	6	7	8	9	10
Topic Title	Waste	Environmental impact assessment	Energy resources	Gene Variation /Noise	Water Pollution	Sea/Wind power	Health /Data	International Cooperation	Water Environment	Climate Change
%	7.3	13.4	10.9	7.4	10.2	9.7	8.4	10.7	12.2	9.8
1	Waste	Environmental impact assessment	Energy	Noise	Pollution	Wind power	Service	Cooperation	Restore	Climate
2	Collection	Evaluation form	Green	Chemistry	Pollution Source	the shore of the sea	statistics	waterand sewage	Wetland	Adaptation
3	Transport	Previous	Greenhouse gas	Gene	Purification	A place of residence	Welfare	The two Koreas	Waterside	Weakness
4	Trash	Consultation	Electric power	Biology	Waste water	Vegetation	Data	exchange	Restoration project	Risk
5	Clean	Collectivity	Emissions right	Engineering	Water quality	the lay of the land	Health	North Korea	Stream	Drought
6	Gas	After	Transaction system	Harmful	Atmosphere	Geology	Information	Northeast Asia	Water intake	Disaster
7	Vehicle	Opinion	Carbon dioxide	Deformation	Sewage	bird	Land Cover Map	South and the North ofKorea	Reservoir	Heat wave
8	Burden	Check	Carbon tax	Human body	Saemangeum	salt flats	Land	Management	receiving in person	Flood waters
9	Diesel	Guide Line	Reduction	Life	Domestic animals	Port	Space	Unification	Water	Temperature
10	Car	Review	Energy source	Food	Animal husbandry	Sea	A research table	Asia	Basin	Change
11	Incineration	Manual	Fossil	Toxicity	Bad smell	Erosion	Function	Agenda	Lower	Rising
12	Occurrence	Procedure	Renewable energy	Extinction	Excreta	water level	Sample	Center	Baekdudaegan	Natural disaster
13	Processing	List	Detergent	Vibration	Sediment	Soil	Disease	Meeting	Aquatic	Typhoon
14	Circularity	Instruction	Coal yard	Microorganism	Heavy metal	Athletics	Database	Private	Jeju	Flood
15	Electricity	Spare	Subsidy	Examination	The upper class	Power generation facility	Variable	Normal	Agricultural water	Warm

3.2.2. Analysis of Naver Environmental News Trends of Each Topic

To analyze the trends of the ten topics extracted from Naver environmental news from 2004 to 2016 in detail, based on the trend analysis results of the KEI research reports, we classified the entire period into three periods (2004 to 2007, 2008 to 2012, and 2013 to 2016). Figure 4 shows the Naver environmental news trends. The three periods in Figure 4 are differentiated by the shadow of the graph background. The analysis results reveal that the amount of articles increased during the entire period except Period2 (2008 to 2012). Chronologically, in 2008, articles on gene variation/noise rapidly increased, and in 2009, articles on sea/wind power, climate change, and environmental impact assessment rapidly increased.

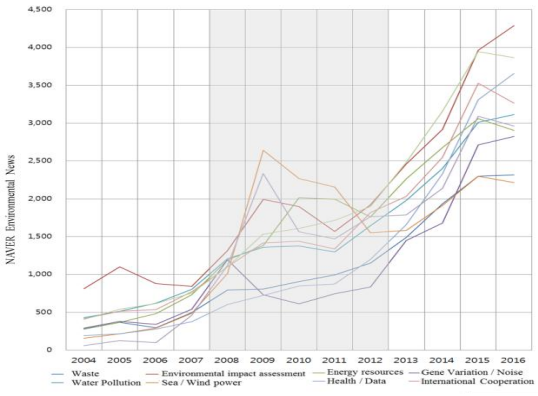


Figure 4:.Naverenvironmental news trend of each topic

3.3. *LDA Analysis of Each Medium*

3.3.1. LDA Comparative Analysis Results of Each Medium

In natural language processing, we compared and analyzed the KEI research reports published at the same time as Naver environmental news collected from 2004 to 2016 (13 years) using LDA-based topic modeling functions. It was found that climate change, waste, environmental impact assessment, energy resources, water pollution, and international cooperation were common topics in Naver environmental news and KEI research reports. When it comes to differences among the media, Naver environmental news articles have gene variation/noise, sea/wind power, and health/data topics, whereas KEI research reports have ecosystem topics. Figure 5 compares the LDA analysis results of individual media on a percentage basis. The green portions refer to common topics in the media, and the orange portions refer to individual media topics.

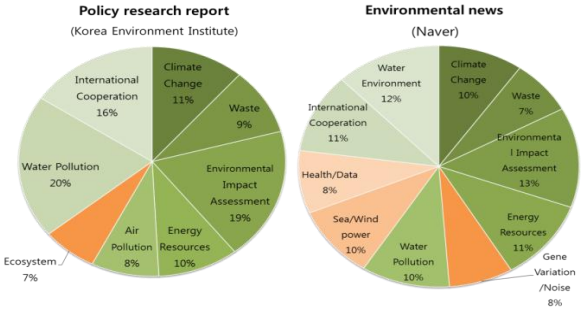


Figure 5:. Comparison of LDA results

3.3.2. Topic Trend Comparative Analysis Results of Each Medium

We compared and analyzed the trends of the common topics (climate change, waste, environmental impact assessment, energy resources, water pollution, and international cooperation) in Naver environmental news and KEI research reports on a percentage basis [Figure 6]. We also examined the trends of the topics (gene variation/noise, sea/wind power, health/data) extracted from Naver environmental news articles, and the topic (ecosystem) extracted from the KEI research reports on a percentage basis [Figure 7]. The trends of each topic were compared and analyzed based on three periods: Period1(2004 to 2007), Period2(2008 to 2012), and Period3(2013 to 2016). Figure 6 shows the analysis results of the six common topics in Naver environmental news articles and KEI research reports. The blue line in the graph indicates Naver environmental news articles, and the red line indicates KEI research reports.

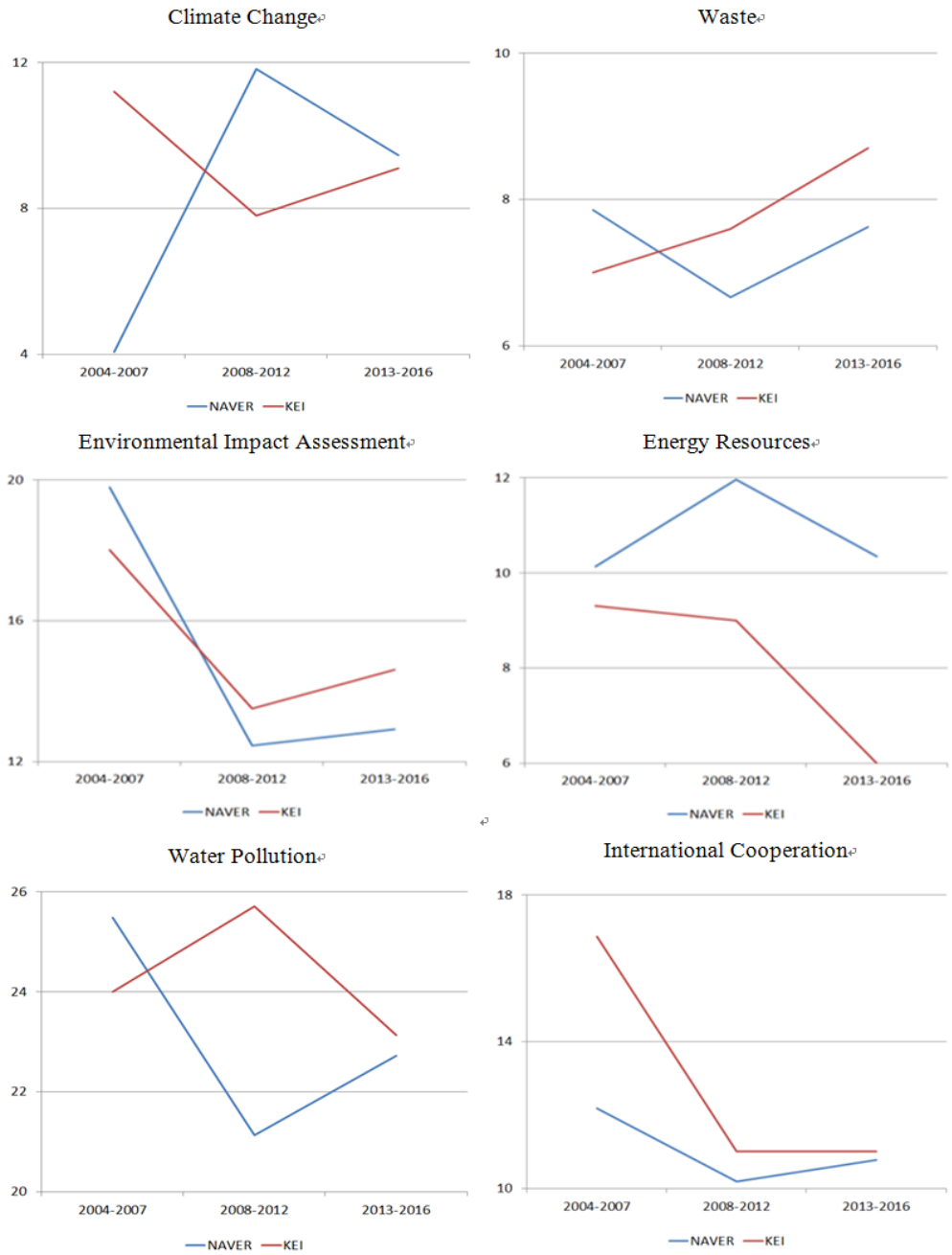


Figure 6:. Common topic trends

The analysis result shows that the topics of environmental impact assessment, energy resources, and international cooperation have similar trends among the media. On Naver, climate change as a topic increased rapidly by 7.77%p,and then decreased by 2.37%p,during Period2, whereas in the KEI research reports, it decreased by 3.40%p and then increased by 1.3%p during the same period. On Naver, waste as a topic decreased by 1.30%p, and then increased by 0.95%p, during Period2, whereas in the KEI research reports, it increased over time. Finally, water pollution as a topic on Naver declined by 4.35%p, and then increased by 1.59%p, during Period2, whereas in the KEI research reports, it increased by 1.71%p, and then decreased by 2.58%p, during Period2. Figure 7 shows the analysis results of the topics of gene variation/noise, sea/wind power, and health/data extracted from Naver environmental news articles, and the topic of ecosystem extracted from KEI research reports. The blue line in the graph indicates Naver environmental news articles, and the red line indicates KEI research reports.

Figure 7:. Distinct topic trends

We analyzed other topics on Naver and discovered that the topic sgene variation/noise fell by 2.50%p, and then increased by 2.31%p, during Period2. On the contrary, sea/wind power as a topic increased rapidly by 7.50%p, and then decreased by 6.20%p, during Period2. The topics health/data rapidly increased 4.29%p during Period3. This demonstrates that, recently, gene variation/noise and health/data related environmental issues have been coming to the fore. These are research topics with growth potential and need to be considered when researchers explore new themes. Finally, in KEI, the topic of the ecosystem increased over time.

4. Conclusion

In this study, we examined the 24-year research trends of the Korea Environment Institute (KEI) using the LDA-based topic clustering technique. We also analyzed whether the KEI research trends corresponded to the social needs for environmental research. Based on the analysis results, we forecasted future environmental research trends and provided reference data that are necessary when researchers develop new research topics. To investigate the social needs for environmental research, we collected environmental news data provided by Naver using the Java HTML parser, jsoup, and conducted an LDA-based topic modeling analysis of the collected data. We then extracted the primary topics of each medium. The LDA-based topic clustering analysis results indicated that, within the KEI research on environmental policies, studies with five topics, namely, energy resources, waste, external cooperation, water environment/environmental impact assessment, and climate change have been carried out over the past 24 years (1993 to 2016). In particular, from 2003 to 2007, there were active studies on climate change, from 2008 to 2012, studies on waste, water environment/environmental impact assessment rapidly increased, and from 2013 to 2016, active studies on waste were conducted.

It was revealed that Naver environmental news articles focused on ten topics waste, environmental impact assessment, energy resources, gene variation/noise, water pollution, sea/wind power, health/data, international cooperation, water environment, and climate change) over the past 13 years (2004 to 2016). Chronologically, there was a rapid increase in articles about gene variation/noise in 2008, and sea/wind power, climate change, and environmental impact assessment in 2009. Overall, articles about all environmental issues increased over the past 13 years except from 2008 to 2012. In addition, we compared and analyzed all data from Naver environmental news articles and KEI research reports from 2004 to 2016 using topic clustering. As a result, climate change, waste, environmental impact assessment, energy resources, water pollution, and international cooperation were commonly deemed important topics. In relation to the differences among the media, on Naver, many environmental news articles regarding health/data and gene variation/noise have been reported in recent years. These are research topics with growth potential, and researchers need to consider them when exploring new research topics.

Based on the analysis results, we confirmed that environmental research trends have followed environmental news trends. Therefore, this study suggests that the exploration of new environmental research topics can be extracted from a periodic analysis of environmental news trends. In addition, we found that the method of identifying the research demand used in this study is a useful way to supplement a traditional qualitative analysis. However, this study failed to use a variety of text data for analysis because only research reports of a specific environmental policy institute and environmental news articles provided by Naver were used, which is a limitation of the study. This implies that further research will need to expand the range of data to social media, scholarly papers, materials published by public institutes, and so on for a deeper analysis. Moreover, analyzing environmental issues from diverse perspectives, including the perspectives of those with demand and suppliers will be helpful in considering environmental issues more closely and from various angles.

Acknowledgment

This study was conducted following the research work 「Big Data Analysis: Application to Environmental Research and Service (GP2017-14)」 and was funded by the Korea Environment Institute.

References

[1] Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77-84.

[2] Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. Handbook of latent semantic analysis, 427(7), 424-440.

[3] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

[4] Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. Handbook of latent semantic analysis, 427(7), 424-440.

[5] Park, S. E. (2015). Analysis of social media contents related to broadcast media using topic modeling. Proceedings of the Korea Intelligent Information Systems Society, 22-22.

[6] Wang, X., & McCallum, A. (2006, August). 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 (pp. 424-433). ACM.

[7] Newman, D. J., & Block, S. (2006). Probabilistic topic decomposition of an eighteenth century American newspaper. Journal of the Association for Information Science and Technology, 57(6), 753-767.

[8] Gerrish, S., & Blei, D. M. (2010, June). A Language-based Approach to Measuring Scholarly Impact. In ICML (Vol. 10, pp. 375-382).

[9] Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In Proceedings of the workshop on interactive language learning, visualization, and interfaces (pp. 63-70).