Journal of Advanced Marine Engineering and Technology, Vol. 44, No. 2, pp. 189~195, 2020
J. Advanced Marine Engineering and Technology (JAMET)
https://doi.org/10.5916/jamet.2020.44.2.189

ISSN 2234-7925 (Print)
ISSN 2234-8352 (Online)
Original Paper

Text mining techniques to identify causes and hazards of ship fire accidents

Byeol Kim[1] · Kwang-Il Hwang[†]

(Received November 14, 2019 ; Revised March 20, 2020 ; Accepted March 23, 2020)

Abstract: SOLAS regulation II-2/17 (MSC/Circ. 1002) allows the use of performance-based fire engineering design to demonstrate that design solutions not complying with traditional prescriptive-based regulations are as safe as equivalent design solutions. However, in the field of fire simulation, which constitutes a large part of performance-based design, there is a paucity of data to set a detailed design of a ship fire scenario. Therefore, the importance of fire hazard elements in designing fire scenarios can vary based on the user's judgments. The purpose of the study is to identify the topics of the causes and hazard of fire accidents for use in the design of fire scenarios. Hence, we used topic modeling of text mining techniques (a text data analysis method) to analyze the written verdict of fire accidents over the past decade. The analysis results are summarized as follows. First, the result of topic modeling indicated that fires due to electrical, mechanical, and carelessness accounted for 69.4%, 17%, and 12.2%, respectively, of the total analyzed fires. Second, fire hazard keywords as per the cause of fire were identified based on the fire cause group obtained through topic modeling. Electrical fire hazard keywords with the highest frequency include "engine room" for fire occurrence location, "electric wire" for flammable material, a "short circuit" for ignition factor, and "flame" for an ignition source.

Keywords: Ship fire, Fire cause, Fire hazard, Topic modeling, Text mining

1. Introduction

1.1 Background and Objectives of the Study

Ships include long-narrow corridors and are composed of steel with high thermal conductivity, which allows fires to spread rapidly on a ship. Additionally, it is not easy to extinguish a fire due to the complex internal structural characteristics of a ship. Furthermore, in the case of initial suppression failure, a ship not only incurs significant damage but also loss of lives **[1][2]**.

Thus, ships are equipped with firefighting and evacuation facilities as per Safety of Life at Sea (SOLAS) rules and regulations to ensure ship fire safety. However, ships in which the regulation is not applicable, the fire safety design can be approved in accordance with SOLAS regulation II-2 / 17 "Alternative design and arrangements". Based on SOLAS regulation II-2 / 17, alternative design and arrangements correspond to fire safety measures that deviate from the prescriptive requirement(s) of SOLAS chapter II. However, they satisfy fire safety objective(s) and functional requirements.

For alternative designs and arrangements, design fire scenarios should be applied to ship designs and computer-based simulations should be used to predict and evaluate fire safety. To develop design fire scenarios that assume a disaster situation, it is necessary to obtain fire hazard data on the sources of ignition, causes of ignition, combustibles, and place of occurrence by fire cause type **[3][4]**.

However, as per statistics of the Korean Maritime Safety Tribunal (KMST), the cause of an accident due to fire is only presented as fire carelessness, namely a defect in the machinery equipment and faulty handling **[5]**. Additionally, 474 keywords related to the cause of the fire accident can be collected from the written verdict summary of the KMST portal **[6]**. However, it is difficult to identify the cause type and the fire hazard elements due to the high number of keywords. Given the lack of historical data, it is difficult for naval architects to set up risk modeling such as fire scenario design.

Therefore, the goal of the present study is to obtain information on the topics of fire accident cause and fire hazard factors from written verdicts of ship fire accidents. Hence, in the study, we

† Corresponding Author (ORCID: http://orcid.org/0000-0003-4850-3558): Professor, Division of Mechanical Engineering, Korea Maritime and Ocean University, 727, Taejong-ro, Yeongdo-gu, Busan 49112, Korea, E-mail: hwangki@kmou.ac.kr, Tel: 051-410-4368

1 Ph. D. Candidate, Department of Refrigeration and Air-Conditioning Engineering, Korea Maritime and Ocean University, E-mail: pooh4762@gmail.com, Tel: 051-410-5030

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Copyright ⓒ The Korean Society of Marine Engineering

applied text mining and topic modeling techniques, which are text data analysis methodology, to extract topics related to causes of fire accidents and to derive keywords. Additionally, the words corresponding to sources of ignition, causes of ignition, combustibles, and place of occurrence by fire accident cause were classified in the order of increasing frequency.

1.2 Literature Review

A few studies examined application of text mining to accident analysis. Kim *et al.* **[7]** used text mining and topic modeling techniques to derive types of accidents from text data contained in chemical plant accident report documents to support the safety managers. Additionally, the relationship between accident type and keywords was represented by the cause-and-effect diagram, and the trend of accident occurrence type with time was summarized. Kang and Suh **[8]** applied text mining to investigate the common causes of the accident, main work type and accident form included in the disaster report documents of the construction industry. Furthermore, self-organizing Map (SOM) algorithm was conducted to visualize the risk factor map. Kang *et al.* **[9]** quantitatively analyzed marine accident data by applying big data techniques to predict the marine accident trend. Thus, they confirmed that it is possible to provide information on preventive measures by grasping objective tendencies for marine incidents that can occur in the future.

However, there is a paucity of studies that apply text mining techniques in the analysis of ship fire written verdicts to derive the topics of fire accident causes and apply them in the design of fire scenarios.

Table 1: Construction of a written verdict of a ship fire accident

Individuals involved in a marine accident	
Text of the judicial decision	
Purport of the claim	
Written reason	
1	Actual fact of marine accident · Name of the ship, Owner of the ship · Port of registration · Gross tonnage, Engine type and power · Individuals involved in the marine accident · Date and location of the marine accident · Location of fire occurrence
2	Cause
3	Action of individuals involved in the marine accident
4	Lessons of accident prevention

2. Scope and Methodology

2.1 Scope

In the study, 147 written verdicts of fire accidents for the last 10 years (2009–2018) were collected from the KMST portal. As shown in **Table 1**, the written verdicts of fire accidents consist of the individual involved in marine accidents, text of a judicial decision, purport of the claim, and written reason. It includes basic information such as the name of the ship, port of registration, gross tonnage, type and power of the engine, date and location of the marine accident, and information on location and cause of fire occurrence, ignition sources, and combustibles.

2.2 Methodology

2.2.1 Text Mining

The written verdict of KMST consists of unstructured data composed of text, and there exists a qualitative analysis method to directly read and arrange the text data although there is a limit to the implementation in the case of high amount of text data. Additionally, the subjective perspective of the individual analyzing the text data can be involved. Thus, it is efficient to use quantitative analysis methods to systematically process text data. Text mining is an unstructured data processing method and refers to a technique that derives useful patterns or relationships and finds meaningful information via applying algorithms or machine learning to unstructured data composed of natural language. It involves data collection, pre-processing, structuring, and frequency or pattern analysis. The text mining process is shown in **Figure 1 [10][11]**.

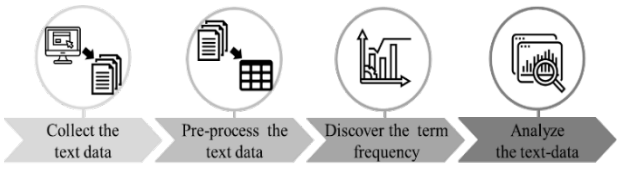

Figure 1: Text mining process

2.2.2 Topic Modeling

The Latent Dirichlet Allocation (LDA) algorithm of the topic modeling method is utilized to classify the types of fire accident causes from the written verdict. Topic modeling is a statistical technique that is used to determine latent topics from the words that make up a large set of documents. It obtains the subject of a document and represents them as clusters of words **[12][13]**. For example, if the cause of fire occurrence is ‘electrical factor’

in a written verdict of a ship fire accident, then it is inferred that words including ‘battery’ and ‘electrical equipment’ appear frequently. Hence, the latent topic can be deduced from the observed words.

The LDA algorithm is a representative method of topic modeling. From the observed variables, such as words in the document, infer hidden variables, such as the topics that make up the documents in each document. The concept of the LDA model is shown in **Figure 2**.

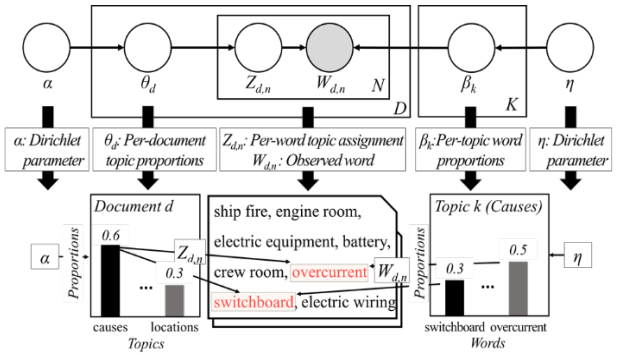

Figure 2: LDA model **[14]**

Specifically, *N* denotes the total number of words collected in the document, *D* denotes the total number of documents, and *K* denotes the number of topics. Node denotes a probability variable and shaded node, $W_{d,n}$ denotes the observed variable in the document as the nth word of document d. When applied to a written verdict of a fire accident, it is observed as words such as 'switchboard' and 'overcurrent'. Furthermore, the hidden variable, $Z_{d,n}$, is the topic of the nth word of document d. **Figure 2** shows that the words ‘switchboard’ and ‘overcurrent’ correspond to the ‘causes’ of the topics. Furthermore, Θ_d denotes a topic distribution included in the d-th document, and β_k is a word distribution constituting the K-th topic. **Figure 2** shows that the distribution of the topic 'Causes' is 60%, and the distribution of the word 'switches' in the topic 'Causes' is 30%. Finally, α and η are parameters that determine θ_d and β_k and can be obtained from the Dirichlet distribution.

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{k} p(\beta_i|\eta) \prod_{d=1}^{D} p(\theta_d|\alpha) \left(\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:k}, z_{d,n}) \right) \quad (1)$$

The LDA probability model is shown in **Equation (1) [11][14][15]**.

2.2.3 Statistical Analysis Tool

In the study, the open-source statistical analysis program R version 3.5.3 is used to analyze ship fire accident written verdict data. We also used the following packages for text data analysis. The KoNLP package is used to remove punctuation and noun extraction from written verdict data, and the tm package is used to generate ‘word × document matrix’ and to apply text mining techniques such as word frequency checking. Furthermore, the topic models package is used for topic modeling.

2.3 Data Pre-Processing

In order to analyze the document of the written verdict composed of text, preprocessing is necessary such that the computer can recognize it. In the study, the data of the written verdict is composed of the corpus. After pre-processing, it is structured into term document matrix (TDM) and statistical analysis is conducted.

The corpus is a collection of documents for analysis and consists of 147 written verdicts of fire accidents. We perform R programming to remove unnecessary information, such as special characters, spaces, punctuation marks, and prepositions of each document in the corpus, for analysis. Additionally, general and non-significant words, such as ‘marine accident’, ‘fire accident’, and ‘adjudication’ in the corpus, are also removed to ensure an effective analysis. Hence, the number of words decreased from 6,788 to 3,401, as shown in **Table 2**.

Table 2: Pre-processing of written verdicts of fire accidents

Text data after pre-processing		
Corpus, representing a collection of text documents		147
Number of terms	Before pre-processing	6,788
	After pre-processing	3,401
Non-/sparse entries		11,934/488,013
Sparsity		98%

Thus, TDM is generated after the process. Specifically, TDM represents the frequency of appearance of a specific word in a specific document in a matrix form in which a document is arranged in a horizontal line and a word is arranged in a vertical line. This involves a matrix of 147 documents and 3,401 words. Additionally, 'Non- / sparse entries' in **Table 2** denote the number of spaces in which frequency information is provided in TDM and the spaces in which frequency information is not provided. Hence, at least 1 frequency is observed in 11,934 cells among 499,947 (= 147 × 3,401) cells although 0 frequency is observed in 488013 cells. Hence, 'Sparsity: 98%' means that

approximately 98% of the cells exhibit a frequency of zero.

Table 3: Sample of term document matrix for the 5 documents

Term \ Document	1	2	3	4	5
engine room	3	31	59	65	29
chief engineer	0	20	35	33	13
Switchboard	0	0	9	2	8
upper deck	14	0	13	6	26
crew room	0	0	2	5	40
owner	0	2	8	6	11
fuel oil	19	17	6	2	1
electric equipment	0	0	8	0	40
wheelhouse	0	5	14	6	5
storage battery	9	0	5	0	0

Table 3 shows part of the TDM that is written for words with high frequency. Therefore, it is possible to understand which words are mainly composed of each document. Documents with many words that can be estimated as a fire location (for e.g., an engine room) can be estimated as a fire accident factor such as a switchboard or electric equipment. Thus, it is possible to infer the relation between words.

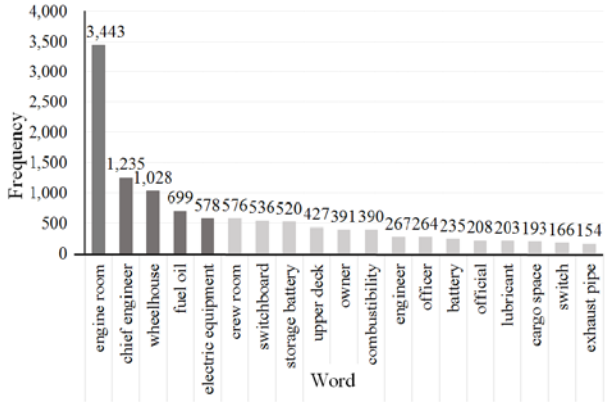

Figure 3: Frequency of words

3. Results

3.1 Word Frequency

To grasp the trend of ship fire accidents in the last 10 years, keywords that are frequently mentioned were extracted from the written verdict data of pre-processed ship fire accidents.

Figure 3 is a graph that shows the frequency of words. The word with the highest frequency (3,443) in the text of the written verdict data is 'engine room', which is inferred as the place of the fire accident. Additionally, other words including ‘fuel oil (699)’, ‘electrical equipment (578)’, ‘switchboard (536)’, ‘storage battery (520)’, and ‘battery (235)’ appear in several documents. The words can be considered as the cause of fire in the engine room. Furthermore, words including ‘wheelhouse (1,028)’, ‘crew room (576)’, ‘upper deck (427)’, and ‘cargo space (193)’ can be viewed as the place of fire or the place of due to fire. Besides, the words ‘chief engineer (1,235)’, ‘engineer (267)’, and ‘officer (264)’ refer to job titles related to marine accidents. **Figure 3** shows the overall trends of the location of fire, causes of fire, and individuals involved in a fire accident over the past decade.

3.2 Topic Analysis of Ship Fire Accident

The LDA algorithm of the topic modeling method was applied to derive a fire accident topic group from the written verdict of a ship fire accident

Prior to applying topic modeling, it is necessary to specify the number of topic groups in advance. There are methods to assume the number of topics based on the subjective judgment of the researcher. However, in the study, the number of subjects was determined by using statistically calculated values as objective indicators **[12][16]**. The number of topics was determined based on the complexity value. The topic group was set up to 50 topics and sampling was repeated 4000 times to select four topic groups with the lowest reduction.

Four topics and seven related keywords were derived as a result of the LDA. Furthermore, the fires caused by electrical, mechanical, and carelessness accounted for 69.4%, 17%, and 12.2%, respectively, of the total analyzed fires.

Table 4: Causes of fire accidents based on topic modeling

Topic Group	Keywords	Frequency
Electrical factors (1)	engine room, crew room, electric equipment, switchboard, electrical outlet, a short circuit, fluorescent light	60(40.8%)
Electrical factors (2)	engine room, engineer, storage battery, battery, overcurrent, source of ignition, electric wiring	42(28.6%)
Mechanical factors	engine room, cylinder, piston, equipment connection, lubricant, long hours, crank	25(17.0%)
Carelessness, arson and other	engine room, cargo space, source of ignition, oil mist, combustibility, gas stove, heater	20(13.6%)
	sum	147(100%)

Table 4 shows the number and frequency of the documents by subject group. The first and second topics correspond to fires caused by electrical factors because the words indicate fires caused by electrical factors such as electrical equipment, switchboard, electrical outlet, short circuit, and overcurrent. The third topic appears to correspond to words, such as engine room, cylinder, piston, connection, and lubricant, thereby indicating that this is a group of fires due to mechanical factors. Finally, the fourth topic can be viewed as a group wherein the cause of fire accidents is due to carelessness because of words, such as gas, stove, heater, and factors other than those mentioned above.

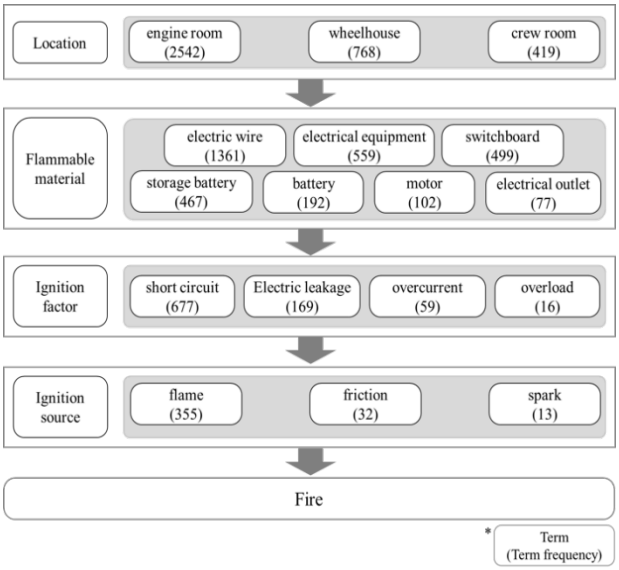

Figure 4: Elements of electrical fire hazard

3.3 Identification of Fire Hazard

Based on the results obtained through the topic analysis, the fire hazard elements for the causes of fire accident are presented in the order of frequency. Fire hazard factors include sources of ignition, causes of ignition, combustibles, and place of occurrence.

As shown in **Figure 4**, an electrical fire can be composed of a scenario where the fire is caused by a flame due to a short circuit of electrical wire and electrical equipment in an engine room. Furthermore, the mechanical fire scenario is shown in **Figure 5**. Mechanical fire scenarios can consist of a fire due to contact with hot components, such as the exhaust pipe of the main engine, and after a fuel oil leak due to damage to the cylinder and piston system. Moreover, the fire scenario can be configured as a fire due to overheating of the heater in the engine room or bedroom. Finally, the carelessness fire scenario consists of a fire that occurs due to a flame near the combustibles in the heater or gas stove in the crew room, as shown in **Figure 6**.

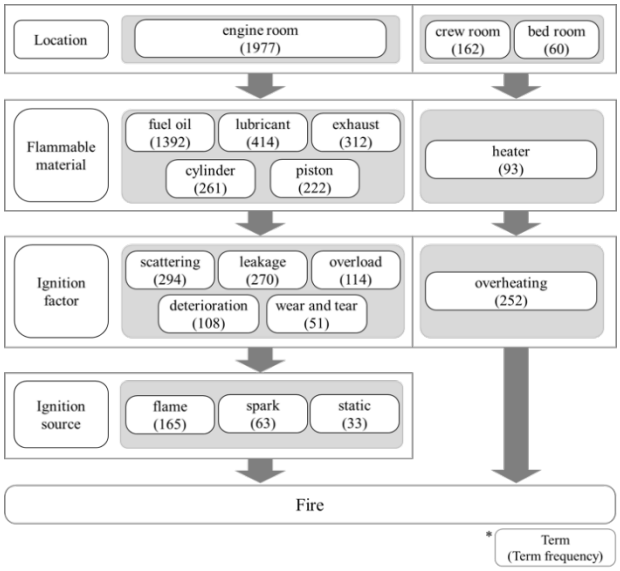

Figure 5: Elements of mechanical fire hazard

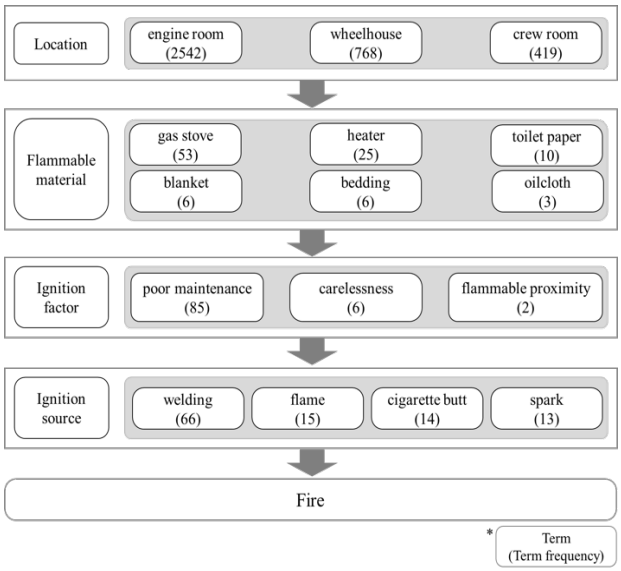

Figure 6: Elements of carelessness fire hazard

4. Conclusion

Based on the Alternative Design and Arrangements for Fire Safety Regulation of SOLAS, performance-based design measures can be applied to evaluate fire safety in the case of ships where it is not possible to apply existing fire safety equipment regulations. As a part of performance-based design, computer-based simulation tools should be used to assign and predict fire scenarios in ship design and safety. However, the

fire safety performance of the ship depends on the user's own judgment given the lack of data and no formal guidelines.

The purpose of the study was to obtain data for use in the design of fire scenarios for evaluating ship fire safety performance. Thus, we applied the text mining technique and LDA algorithm to written verdicts of ship fire accidents in the last 10 years to derive the causes of fire accidents and fire hazard factors. The major research results are summarized as follows:

First, the topic analysis on the causes of ship fire accidents indicated that the main causes were electrical factors (69.4%), mechanical factors (17%), carelessness, arson, and other factors (13.6%).

Second, the location of the fire occurrence, flammable material, ignition factor, and ignition source were obtained as fire hazard elements. Electrical fires were observed as the highest in the engine room due to short circuits in the electrical wires. Additionally, fuel oil fire was the most frequent fire due to mechanical factors and scattering acted as an ignition factor. Furthermore, carelessness fires were determined as due to the proximity of flammables in gas stoves or heaters and fires due to welding embers being transferred to flammables during welding.

The text mining and topic modeling techniques used in the study can effectively identify the causes of ship fire accidents and fire hazard elements included in the written verdict. In the future, it is necessary to examine how the results of the study can be used in prescribing design fire. Additionally, we plan to apply the fire risk factors obtained in the study to actual fire simulation scenarios and use the same in fire safety performance evaluation.

Author Contributions

Conceptualization, B. Kim and K. I. Hwang; methodology, B. Kim; Software, B. Kim; Formal Analysis, B. Kim; Investigation, B. Kim; Resources, B. Kim; Data curation B. Kim; Writing-Original Draft Preparation, B. Kim; Writing-Review & Editing, B. Kim and K. I. Hwang; Visualization, B. Kim; Supervision, K. I. Hwang; Project Administration, K. I. Hwang.

References

[1] K. I. Hwang, I. S. Cho, G. H. Yun, and B. Kim, "A comparison of the trainees' evacuation characteristics according to the indoor smoke-fullfill during the safety training on ship," Journal of the Korean Society of Marine Environment & Safety, vol. 24, no. 4, pp. 422-429, 2018.

[2] B. Kim and K. I. Hwang, "Smoke exhaust performance prediction according to air supply and exhaust conditions for shipboard fires from a human safety point of view," Journal of the Korean Society of Marine Environment & Safety, vol. 22, no. 7, pp. 782-790, 2016.

[3] H. J. Kang, J. Choi, D. K. Lee, and B. J. Park, "A framework for using computational fire simulations in the early phases of ship design," Ocean Engineering, vol. 129, pp. 335-342, 2017. Available: https://doi.org/10.1016/j.oceaneng.2016.11.018

[4] International Maritime Organization (IMO), "Guidelines on alternative design and arrangements for fire safety," UK, MSC/Circ. 1002, 26 June 2001.

[5] KMST, Korea Maritime Safety Tribunal, Statistics for the marine accidents, https://www.kmst.go.kr/kmst/statistics/annualReport/selectAnnualReportList.do (in Korean), Accessed October 31, 2019.

[6] KMST, Korea Maritime Safety Tribunal, Written verdicts of Korea maritime safety tribunal, https://www.kmst.go.kr/kmst/verdict/writtenVerdict/selectWrittenVerdict.do (in Korean). Accessed October 31, 2019.

[7] B. S. Kim, S. R. Chang, and Y. Y. Suh, "Text analytics for classifying types of accident occurrence using accident report documents," Journal of the Korean Society of Safety, vol. 33, no. 3, pp. 58-64, 2018 (in Korean).

[8] S. S. Kang and Y. Y. Suh, "On the development of risk factor map for accident analysis using textmining and self-organizing Map (SOM) algorithms," Journal of the Korean Society of Safety, vol. 33, no. 6, pp. 77-84, 2018 (in Korean).

[9] S. Y. Kang, K. S. Kim, H. B. Kim, and B. S. Rho, "An analysis of causes of marine incidents at sea using big data technique," Journal of the Korean Society of Marine Environment & Safety, vol. 24, no. 4, pp. 408-414, 2018 (in Korean).

[10] A. Hotho *et al.*, A brief survey of text mining, Journal for Computational Linguistics and Language Technology, vol. 20, no. 1, pp. 19-62, 2005.

[11] Y. M. Baek, Text Mining Using R, Korea, Han-ul Academy, 2017 (in Korean).

[12] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” Journal of Machine Learning Research, vol. 3, pp. 993-1022, 2003.

[13] M.-K. Kim, Y. Lee. and C.-H. Han, “Analysis of consulting research trends using topic modeling,” Journal of the Society of Korea Industrial and Systems Engineering, vol. 40, no. 4, pp. 46-54, 2017 (in Korean).

[14] D. M. Blei, “Probabilistic topic models,” Communications of the ACM, vol. 55, no. 4, pp. 77-84, 2012.

[15] J. H. Park and M. Song, “A study on the research trends in library & information science in Korea using topic modeling,” Journal of the Korean Society for Information Management, vol. 30, no. 1, pp. 7-32, 2013 (in Korean).

[16] B. Grün and K. Hornik, “Topicsmodels: An R package for fitting topic models,” Journal of Statistical Software, vol. 40, no. 13, pp. 1-30, 2011. Available: http://www.jstatsoft.org/v40/i13.