Analysis of Safeguardability Assessment Parameters Using Text Mining

Geunhee Kim and Seungmin Lee*
Korea Institute of Nuclear Nonproliferation and Control (KINAC), 1418 Yuseong-daero, Yeseong-gu, Daejeon, 31030
**Corresponding author: seungmin@kinac.re.kr*

1. Introduction

The IAEA encourages a safeguards-by-design (SBD) approach that considers safeguards from the design stage for new nuclear facilities [1]. In the Republic of Korea, new nuclear facilities such as small reactors and interim spent fuel storage facilities are expected, and safeguardability assessment technologies and legal frameworks are insufficient. To ensure the implementation of safeguards for these new nuclear facilities, regulatory agencies must review and evaluate whether safeguards considerations are adequately reflected in the design information provided by the designers. The development of safeguardability assessment factors and tools for supporting regulatory activities is currently underway [2].

In this study, we used text mining techniques to verify whether safeguardability assessment factors could be appropriately selected and applied to new facilities. Text mining is a powerful technique for analyzing and extracting useful information from large amounts of unstructured text data [3]. Using this technique, we can explore and extract objective information without bias, and save time and resources. In particular, the term frequency-inverse document frequency (TF-IDF) method is a statistical method used to evaluate how often a word or term appears in a document and how unique it is in a collection of documents. To develop guidelines for applying safeguards in nuclear facilities, we aimed to identify the most relevant safeguardability assessment factors based on their frequency and uniqueness in a corpus of safety manuals, regulations, reports, and related documents using TF-IDF [4].

2. Safeguardability assessment parameters

In the ongoing research project, we first analyzed previous studies related to nuclear proliferation resistance and safeguardability assessment, and derived general safeguardability assessment parameters. To verify this, we used the Delphi technique and the analytic hierarchy process to reduce the number of parameters to 19, as summarized in Table I. The assessment parameters were divided into three elements: Design information verification (DIV), nuclear material accountability, and containment and surveillance. Each higher element was assigned six, seven, or six subelements (19 safeguardability assessment parameters). This can be appropriately used as a tool to evaluate safeguardability. In this study, we aim to compare the importance of these parameters through text mining techniques.

Table Ⅰ: Safeguardability assessment parameters

Higher elements	Sub-elements (Safeguardability assessment parameters)
Design information verification	1. Completion of design information
	2. Access of inspectors to essential equipment in the nuclear facilities
	3. Access of inspectors to the entire nuclear facility during the construction or operation process
	4. Minimizing radioactivity levels during DIV
	5. Management of documentation related to safety protocols such as design information
	6. Including summary of changes in the design information in a timely manner
Nuclear materials accountancy	7. Use of nuclear materials verification equipment (NDA, DA)
	8. Independent storage location dedicated for nuclear material verification equipment
	9. Lighting and space for the nuclear material storage space
	10. Able to identify the storage location of nuclear materials in the storage space
	11. Able to attach ID tags on the nuclear material and identify them
	12. Able to dismantle or reconstruct nuclear material items according to their types
	13. Calibration of nuclear material measuring instruments
Containment and Surveillance	14. Uninterrupted power supply for containment device
	15. Access of inspectors to the containment structures (such as walls) of the nuclear facility
	16. Standardization of access path and frequency
	17. Uninterrupted power supply for surveillance equipment
	18. Inclusion of sealing and surveillance equipment when designing nuclear facilities
	19. Communication facility dedicated for safety protocol

3. Methods

Text mining is a technology that finds valuable and meaningful information from unstructured natural language data by extracting object names, patterns, or word-sentence relationships. Current analysis methods include morphological analysis, vector space modeling, semantic network analysis, and text-sentiment analysis. In this study, we performed frequency analysis, a representative method of text mining, using term

frequency and inverse document frequency to extract keywords from documents.

3.1 TF (Term Frequency)

Frequency analysis is the most intuitive and widely used text mining technique. By determining the frequency of words in a document, we can identify and visualize the main keywords. TF analysis is a simple method of identifying frequently used words by counting the number of times each word appears in a document. Using the countvectorizer class in the Python library scikit-learn, we can count the occurrence frequency of words and vectorize them. Furthermore, based on the obtained term frequency, we can visualize words in proportion to their size using the Python library word cloud [5].

3.2 TF-IDF (Term Frequency - Inverse Document Frequency)

TF-IDF is a statistical measure that indicates the importance of a word in a specific document, given a collection of documents. Because frequently used universal words commonly appear across multiple documents, the results obtained through the TF analysis may show that topic words with significant importance in a specific document have a low frequency. Therefore, to determine the relative importance of a topic word within a specific document, we considered the IDF value by weighting it based on the ratio of documents containing the topic word to the total number of documents. In other words, if a word appears frequently in the entire collection of documents, its TF value is high, and if it appears less frequently in the entire document compared with a specific one, its IDF value is high. Therefore, to understand the importance of a word within a given document, we must consider both TF and IDF. The calculation method for the TF-IDF is as follows [6]:

TF-IDF (Term x within document y) :

$$\omega_{x,y} = \mathrm{t}f_{x,y} \times \log\frac{N}{df_x}$$

$\mathrm{t}f_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

4. Results

Text collection, preprocessing, and analysis were performed for text mining. First, we collected and extracted texts, including the safeguards implementation practice guide, Nuclear Power (NP) technical reports of the IAEA, and research documents related to safety assessment factors in an electronic document format. Data preprocessing was performed to remove unnecessary words and improve text quality. In this study, we performed tokenization, cleansing, normalization, and stop word removal for text preprocessing. Based on this, we applied text mining techniques to analyze the data. In this section, we describe the results obtained using the TF and TF-IDF methods and the visualization materials.

4.1 TF

The frequency of the words was measured in 26 documents, and the results are summarized in Table II. Universal keywords representing the nuclear industry, such as “nuclear,” “IAEA,” “fuel,” and “energy,” were identified.

Among the results, keywords such as “safeguards,” “design,” “material,” “facility,” “information,” and “proliferation” can be considered as keywords related to the main topic of this study, which is “safeguards by design (SBD) of new nuclear facilities.” In particular, it is related to DIV and nuclear material accountancy, which are the higher elements for safeguardability assessment. The word cloud visualizes the analysis, with larger words indicating a higher frequency, which helps to understand the content and topics of the documents intuitively. Through frequency analysis, we confirmed that not only universal keywords related to nuclear facilities but also words related to our research topic were among the top results.

Table II: Result of Term frequency (TF)

R	Term	f	R	Term	f
1	nuclear	6058	11	state	1567
2	safeguards	5853	12	proliferation	1487
3	IAEA	4459	13	international	1458
4	design	3998	14	system	1452
5	material	3692	15	requirements	1361
6	facility	3424	16	equipment	1343
7	fuel	2417	17	facilities	1333
8	process	1811	18	activities	1137
9	energy	1780	19	systems	1110
10	information	1580	20	measures	1056

*f : frequency; R: rank

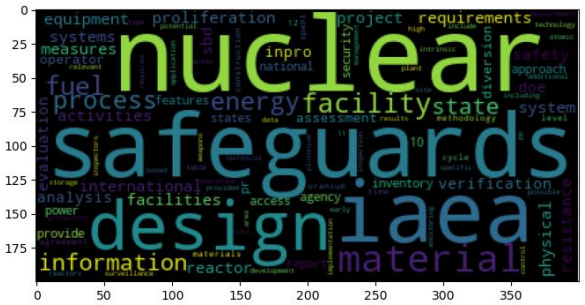

Fig. 1. Wordcloud of TF results

4.2 TF-IDF

We compared the keywords extracted based on the TF-IDF analysis with the results of the TF analysis. As summarized in Table III, the top eight words were similar in both analyses and did not exhibit clear differences. However, from the 9th word, words such as "SBD," "resistance," "project," and "INPRO" appeared as top keywords in TF-IDF analysis. This indicates that keywords related to safeguard assessment in the design phase are more important. It also implied that the research methodology of "INPRO" or "FSA project" could be a significant role in determining assessment parameters.

The importance of each assessment parameter derived from the ongoing research project was compared using the TF-IDF scores. First, we compared each higher element by adding the scores of the words that constituted them. For instance, the "DIV" element has a score of 6.7244, which is the sum of 5.0748 for "design," 1.1511 for "information," and 0.4985 for "verification." Among three higher elements, "nuclear materials accountancy (NMA)" had the highest importance (10.2182), followed by DIV (6.7244) and CS (0.3097).

Next, we compared the subelements. First, the TF-IDF scores of all the words constituting each sub-element, such as higher elements, were added to rank them. This is summarized in Table IV, and the number of parameters is sequentially assigned to the identifiable parameters in Section 2. The top five factors were factors 18, 5, 3, 4, and 1, [1)]which simply reflected the scores of all the words that constitute each parameter. Therefore, we need to [2)]exclude non-essential parts of parameters, such as "during ~" and "such as ~," to reflect the meaning more accurately. However, this method still had the drawback of not reflecting the importance of higher elements.

Therefore, we [3)]added the scores of each sub-element to the scores of the higher elements and compared them, and as a result, factors 2, 3, 7, 18, and 10 exhibited higher importance. Access to equipment or facilities by inspectors and the use of nuclear material verification equipment are crucial factors in protecting nuclear facilities and preventing materials from being diverted. In addition, identification of the storage location of materials allows for effective and accurate monitoring and control of nuclear materials. Sealing and surveillance equipment are important to detect the diversion and misuse of nuclear material.

Table Ⅲ: Result of TF and TF-IDF analysis

R	TF		R	TF-IDF	
	Term	f		Term	score
1	nuclear	6058	1	safeguards	6.5193
2	safeguards	5853	2	nuclear	6.2505
3	IAEA	4459	3	design	5.0748
4	design	3998	4	IAEA	4.5808
5	material	3692	5	facility	3.8562
6	facility	3424	6	material	3.6106
7	fuel	2417	7	fuel	1.8522
8	process	1811	8	process	1.7354
9	energy	1780	9	SBD	1.5088
10	information	1580	10	proliferation	1.2202
11	state	1567	11	information	1.1511
12	proliferation	1487	12	requirements	1.1075
13	international	1458	13	energy	1.0924
14	system	1452	14	international	1.0311
15	requirements	1361	15	resistance	0.9562
16	equipment	1343	16	project	0.9335
17	facilities	1333	17	state	0.9017
18	activities	1137	18	equipment	0.8461
19	systems	1110	19	inpro	0.7517
20	measures	1056	20	reactor	0.7491

*f : frequency; R: rank.

Table Ⅳ: Top five parameters by calculation of TF-IDF scores

R	1)		2)		3)	
	No.	score	No.	score	No.	score
1	18	10.7276	18	10.7276	2	12.5368
2	5	7.3004	1	6.2259	3	11.9757
3	3	6.9865	2	5.8123	7	11.5628
4	4	6.8296	19	5.2694	18	10.7949
5	1	6.2259	3	5.2512	10	10.4330

*R: rank; No.: number of parameters.
1): simply all the words of sub-element parameters
2): exclude non-essential parts of sub-element parameters
3): higher element score + sub-element score by 2) calculation

5. Conclusions

In this study, we used text mining techniques to develop assessment parameters for applying safeguards by design (SBD) of new nuclear facilities. We compared the TF and TF-IDF analyses of related documents to identify the most critical assessment parameters and elements. In particular, the higher element "NMA" and sub-element No.7 and 19 obtained high importance, which was confirmed to be consistent with other expert evaluations. In future research, it will be necessary to collect more information and apply other text-mining techniques or examine changes in the safeguardability assessment over time.

ACKNOWLEDGEMENT

This work was supported by the Nuclear Safety Research Program through the Korea Foundation of Nuclear Safety (KoFONS), using financial resources granted by the Nuclear Safety and Security Commission (NSSC) of the Republic of Korea (No. 2106018).

REFERENCES

[1] International Atomic Energy Agency, International safeguards in nuclear facility design and construction, IAEA Nuclear Energy Series No.NP-T-2.8, 2013

[2] B.Y. Kim, S.-K. Ahn, H.-D. Kim, and D.-Y. Song, A Study of Safeguardability Evaluation Approach for implementation of SBD, Transactions of the Korean Nuclear Society Autumn Meeting, 2022
[3] Tan, A. H., Text mining: The state of the art and the challenges, Proceedings of the PAKDD Workshop on Knowledge Discovery from Advanced Databases, 8, pp. 65-70, 1999
[4] Havrlant, L., & Kreinovich, V., A simple probabilistic explanation of term frequency-inverse document frequency(TF-IDF) heuristic (and variations motivated by this explanation), International Journal of General Systems, 46(1), pp. 27-36, 2017
[5] Florian H., Steffen L., Simon L., Thomas E., Word Cloud Explorer: Text Analytics based on Word Clouds, 47th Hawaii International Conference on System Science, 2014
[6] Ramos, J., Using TF-IDF to determine word relevance in document queries, Proceedings of the first instructional conference on machine learning, 242(1), pp. 29-48, 2003