A TF-IDF based Retrieval Method for Entry of Actions of LCO Cases

Nokyu Seong*, Sangwon Oh
KHNP Central Research Institute, 70, 1312 beon-gil, Yuseong-daero, Yuseong-gu, Daejeon
**Corresponding author: nokyuseong@khnp.co.kr*

Keywords : Technical Specifications, Limiting Conditions for Operation, TF-IDF, Natural Language Processing

1. Introduction

Limiting Conditions for Operation (LCO) are critical elements of the Technical Specifications (TS) in nuclear power plants and must be continuously monitored in real time to ensure safe plant operation. When an LCO is not met, corresponding required actions must be executed within a specified completion time, and the entire process must be properly documented [1]. In domestic nuclear power plants, Entry into Actions of LCO (EAL) cases are managed in a digitalized manner through the introduction of Enterprise Resource Planning (ERP).

However, current systems rely primarily on lexical-based search using keyword matching, which leads to low retrieval precision and difficulty in extracting meaningful information despite the abundance of data. Furthermore, the unstructured of historical LCO data limits the effective utilization of accumulated operating experience.

To address these challenges, this study designs a retrieval prototype employing TF-IDF (Term Frequency-Inverse Document Frequency) to optimize the retrieval of unstructured LCO cases and validates its effectiveness. The proposed methodology is expected to enhance operators' decision-making capabilities and establish consistent LCO application principles across different plant sites and reactor types, thereby improving the safety and operational efficiency of nuclear power plants.

2. TF-IDF & Vector Space Model

To address the limitations of keyword-based retrieval discussed in previous section, this chapter outlines the fundamental concepts of the vector space model and the TF-IDF weighting method, which are widely used in document similarity measurement and information retrieval.

2.1. Term Frequency & Inverse Document Frequency

TF-IDF represents documents as weighted vectors by reflecting both the importance of a term with in a document and its discriminative power across a document collection. The term frequency measures how frequently a term appears in a given document, while the inverse document frequency reduces the weight of terms that appear frequently across many documents. The TF-IDF weight is calculated as the product of TF and IDF, enabling more informative terms to have higher influence in document representation and similarity computation [2]. In notifications related to LCO, terms such as 'LCO', 'EDG', and 'HVAC' appear with different frequencies across the document corpus. The term 'LCO' is included in a large proportion of documents, whereas equipment-related terms such as 'EDG' and 'HAVC' occur in a more limited subset of cases. As a result, the TF-IDF weights of these terms vary according to their document frequencies. Table I presents an example of TF-IDF weighting results, showing how commonly occurring terms receive lower weights, while more discriminative terms are assigned higher wights.

Table I: Example of TF-IDF weights

	LCO	EDG	HVAC
Weight	1.07	2.88	4.09

2.2. Vector Space Model

The vector space model represents documents and queries as vectors in a common multidimensional space and serves as a fundamental framework for information retrieval tasks such as document ranking, classification, and clustering. In this model, each document is expressed as vector whose elements correspond to weighted terms, and a query is represented in the same vector space. When both documents and queries are represented as vectors, their similarity can be quantified using cosine similarity, which is computed based on the normalized dot product of the two vectors. By ranking documents according to their similarity scores, the vector space model enables efficient retrieval of LCO cases that are most relevant to a given query [3].

3. Results and Discussion

This study analyzes approximately 18,000 LCO notification records accumulated between 2004 and 2024, using the notification titles as the primary textual data for similarity-based retrieval. Each notification title is treated as an individual document, and user queries are compared against these documents to identify similar LCO cases. For document tokenization, Korean characters, English letters and numerical values were included, along with selected special symbols such as '.', '/', '-', '(', ')', '[' and ']'. Tokenization was performed based on whitespace without applying a morphological analyzer.

3.1. Results

Based on a user survey, the query patterns used for searching EAL cases can be classified into three representative categories. The Query type 1 consists of queries combining plant location, system or/and equipment and condition. The Query type 2 includes system or equipment, parameter, and condition. The Query type 3 is composed of plant type, LCO item and EAL cases. The proposed retrieval approach was evaluated by applying representative queries from each category. For each representative query type, the top ten EAL cases with the highest similarity scores were retrieved and ranked in similarity. Although actual user queries and EAL cases were collected in Korean, they are replaced in this paper with semantically equivalent English expressions for clarity and consistency.

Table II presents the retrieval results for Query type 1 which combines plant location, system, equipment, and condition. As shown in the table, the similarity scores of the top ten retrieved cases range from 0.45 to 0.56. For the highest ranked results, three query tokens are matched, and the relatively short document length contributes to higher similarity scores. In contrast, the documents ranked 8th to 10th also contain three matching tokens; however, their longer document lengths result in slightly lower similarity scores. As a result, these documents exhibit lower similarity values compared to documents with only two matching tokens but shorter overall lengths.

Table II: 'MCR HVAC FAN Inoperable'

Rank	Title	Sim.
1	MCR HVAC "B" Train FAN 01CB Inspection	0.56
2	MCR HVAC "B" Inoperable	0.48
3	MCR HVAC "A" Inoperable	0.48
4	MCR HVAC "A" Inoperable	0.48
5	MCR HVAC "B" Inoperable	0.48
6	MCR HVAC "B" Inoperable	0.48
7	MCR HVAC "B" Inoperable	0.48
8	[LCO]U4 MCR HVAC Return FAN Low Flow Trip	0.47
9	MCR HVAC B Train EMER FAN(AH-03B) Inspection	0.45
10	MCR HVAC A Train RTNR FAN 02CA Inspection	0.44

Table III shows the retrieval results for Query type 2, focusing on queries composed of equipment or system, parameter and condition. For the second query type, the similarity scores show high values ranging from 0.89 to 0.82. Although all five query tokens are present in some retrieved documents, certain cases exhibit lower similarity scores than the top-ranked document which has a shorter document length. This indicates that document size of length affects the similarity calculation, resulting in lower similarity scores for longer documents despite full token matching.

Table III: 'RCS Cold Leg Temperature Exceeded'

Rank	Title	Sim.
1	RCS Cold Leg Temperature	0.89
2	[LCO] RCS Cold Leg Temperature Limit Exceeded	0.89
3	[LCO] U3 RCS Cold Leg Temperature Limit Exceeded	0.82
4	[LCO] U3 RCS Cold Leg Temperature Limit Exceeded	0.82
5	[LCO] U3 RCS Cold Leg Temperature Limit Exceeded	0.82
6	[LCO] U4 RCS Cold Leg Temperature Limit Exceeded	0.82
7	[LCO] U4 RCS Cold Leg Temperature Limit Exceeded	0.82
8	[LCO] U4 RCS Cold Leg Temperature Limit Exceeded	0.82
9	SK3 LCO 3.4.1 RCS Cold Leg Temperature	0.73
10	EAL related RCS Cold Leg Temperature Limit Exceeded	0.73

Table IV describes the retrieval results for Query type 3 which addresses plant type, LCO item and EAL cases. For Table IV, the similarity scores range from 0.40 to 0.21. Consistent with the results in Table III, the similarity values are influenced by document length, confirming that document length affects the similarity calculation.

Table IV: 'OPR 3.1.7 EAL Cases'

Rank	Title	Sim.
1	LCO 3.1.7 EAL	0.40
2	SK4 LCO 3.1.7 Regulating CEA Insertion Limits EAL	0.23
3	SK4 LCO 3.1.7 Regulating CEA Insertion Limits EAL	0.23
4	SK4 LCO 3.1.7 Regulating CEA Insertion Limits EAL	0.23
5	SK4 LCO 3.1.7 Regulating CEA Insertion Limits EAL	0.23
6	[LCO] Y6 3.1.7 Regulating CEA Insertion Limits EAL	0.22
7	[LCO] Y5 3.1.7 Regulating CEA Insertion Limits EAL	0.22
8	[LCO] Y5 3.1.7 Regulating CEA Insertion Limits EAL	0.22
9	[LCO] Y6 3.1.7 Regulating CEA Insertion Limits EAL	0.22
10	[LCO] TS 3.1.7 Regulating CEA Insertion Limits EAL	0.21

3.2. Discussion

The results demonstrate that the TF-IDF based retrieval methods is effective in identifying EAL cases relevant to diverse user query patterns. By assigning higher weights to discriminative terms, the proposed approach improves the ranking of meaningful EAL cases compared to simple keyword-based retrieval. These observations indicate that, given the short length of

notification titles, document length plays a significant role in similarity computation because the maximum document length is only 12.

Furthermore, retrieval performance can be improved through the integration of additional techniques beyond TF-IDF weighting alone. In particular, ontology data structuring and filtering, such as a plant type, system classification, equipment category, and operational status, can enhance retrieval efficiency. A hybrid approach that combines TF-IDF based similarity ranking with structured filtering is therefore expected to provide more accurate and practical search results.

In addition, the use of domain-specific resources including a nuclear terminology dictionary, synonym dictionaries, and stop-word lists is expected to further improve retrieval performance. These improvements, together with the development of a structured database for systematic LCO knowledge management, will be addressed in future work.

4. Conclusion

This paper proposed a TF-IDF based retrieval method for searching EAL cases in order to overcome the limitations of conventional keyword-based search approaches. User queries collected from operators were categorized into three representative types, and retrieval results for each type were presented to demonstrate the effectiveness of the proposed method. The results confirm that the TF-IDF based similarity retrieval can effectively identify relevant EAL corresponding to operators' queries. Furthermore, it is expected that performance can be significantly improved by structuring the data using an ontology-based approach and presenting the results in a hybrid manner that combines similarity-based ranking with structured filtering. Through the proposed method, operators are expected to more easily retrieve relevant past cases when determining whether LCO is met or not met, thereby supporting more efficient and informed decision-making.

REFERENCES

[1] U.S. Nuclear Regulatory Commission (NRC), Standard Technical Specifications Combustion Engineering Plants Revision 4.0 Volume 1, Specifications NUREG-1432- U.S. Nuclear Regulatory Commission (NRC), 1, 2012

[2] C. D. Manning, P. Raghavan, H, Schütze, Introduction to information retrieval, Natural Language Engineering, 2010

[3] N. K. Seong, J. H. Lee, J. B. Lee, P. H. Seong, Retrieval methodology for NPP LCO cases based on domain specific NLP, Nuclear Engineering and Technology, Vol. 55, p. 421-431, 2023