Conceptual Analysis of Integrated Database and Semantic Search Engine of Nuclear Non-proliferation Data

Byoungchan Han*, Byeongmun Ahn, Tongkyu Park, Sung-Kyun Zee
FNC Technology Co., Ltd., 32F, 13 Heungdeok 1-ro, Giheung-gu, Yongin-si, Gyeonggi-do, Korea
**Corresponding author: bchan007@fnctech.com*

1. Introduction

Republic of Korea has joined the international nuclear non-proliferation regime such as Treaty on the Non-Proliferation of Nuclear Weapons (NPT) for peaceful use of nuclear power and prohibiting proliferation of the nuclear weapons. In addition, Korean government signed Safeguards Agreements with the International Atomic Energy Agency (IAEA) to improve international reliability and transparency in nuclear power.

In order to comply with the international nuclear non-proliferation regime, the regulatory agency is carrying out export control on nuclear-related strategic items specified in the Nuclear Safety Act [1]. Moreover, databases containing various types of non-proliferation data are established and operated for this purpose. Meanwhile, the number of imports and exports of nuclear-related strategic items is expected to increase in the future, as domestic nuclear technology improves and international cooperation increases. As a result, there has been an increase in demand for integrated databases and semantic search engines in order to fulfill export control operations efficiently and effectively.

2. Preliminaries

2.1. Data warehouse

Data warehouse is a kind of data storage system used for data analysis. [2] It is a centralized repository that stores many kinds of distributed data in single place. The difference between conventional relational database and data warehouse is that relational databases focus on data transactions such as creating, reading, updating, deleting (CRUD) with maintaining data integrity, while data warehouses focus on rapid, high-dimensional analysis of data. In addition, operational databases are not suitable for data analysis, because such analysis can cause significant loads as they perform data transactions in real-time. This is the reason why companies often build data warehouses.

As stated above, data warehouses are used to obtain business intelligence through data analysis. For this process, which is called online analytical processing (OLTP), most data warehouses are built in column-oriented model for data processing. By storing data in columns, column-oriented databases, as opposed to row-oriented databases, can access data more quickly without scanning and discarding unwanted row data. Also, the data warehouse is designed in a multidimensional cube modeling method to analyze the same data from various perspectives. Star schema method and the snowflake schema method shown in Fig. 1 are typical examples of this model.

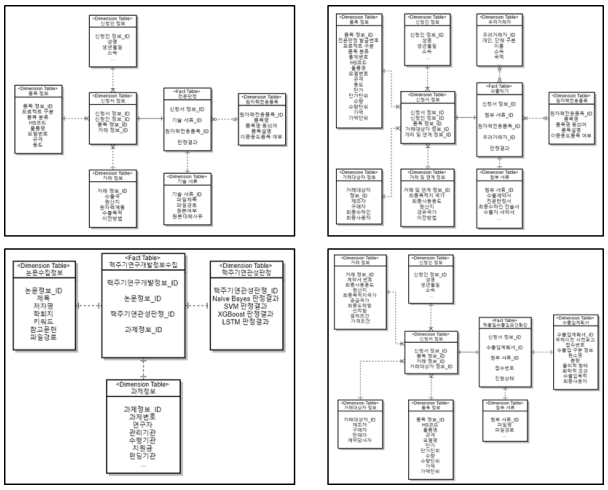

Fig. 1. Examples of data warehouse entity-relationship diagram based on star schema and snowflake schema methods.

2.2. Semantic search

Semantic search is a search method in which a computer interprets the overall meaning of the search query and returns the results to the user. Unlike lexical search, which only finds for literal matches to the query, semantic search improves its accuracy by understanding the user's intent as well as the contextual meaning of terms. To understand the user's intention, previously developed semantic search engines map the relationship between the contextual meaning of the search word and the data stored, and build ontology to return data that are highly related to the keywords in the search query.

Deep learning techniques are frequently used in the latest semantic search engines. This method employs word embedding techniques to convert a set of words into the real-valued vectors that encode the meaning of the words, with the expectation that closer vectors will have more similar meanings. Word embedding techniques instruct computers to learn the meanings of words and represent them as vectors in n-dimensional space. Examples include Word2Vec, Glove, and BERT which recently demonstrated superior performance in natural language processing.

2.2. BERT

The BERT model refers to Bidirectional Encoder Representations from Transformers which designed for language representation. [3] As it considers the context of both directions, it is able to grasp the meaning of words more naturally. Furthermore, traditional embedding models must be trained from scratch for each task; however, BERT is pre-trained from a large number of Wikipedia and BooksCorpus words and can be used for transfer learning. BERT can be used for a wide range of NLP tasks, including question and answer, machine translation, and topic search. It has achieved cutting-edge results in 11 NLP task benchmarks.

3. Conceptual Analysis

For the analysis of various data, an integrated database is created in the form of a data warehouse. The data warehouse is divided into data marts based on the export control tasks, and each data mart provides the analysis results requested by the users. Data marts are created for each independent export control task, such as classifying materials, managing strategic item exports, and verifying nuclear material import/export conditions. The overall architecture of the integrated database is depicted in Fig. 2.

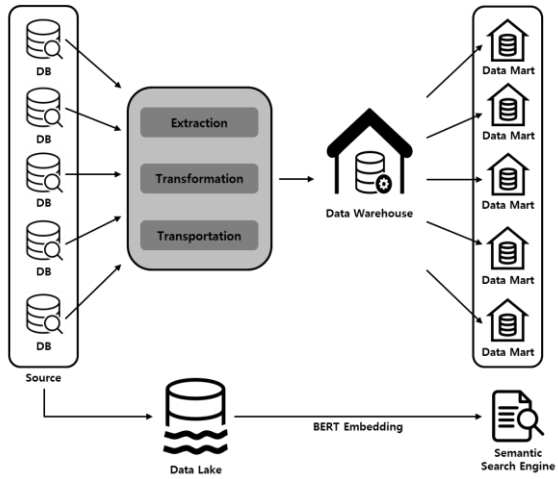

Fig. 2. Architecture of the overall integrated database.

In addition, the integrated database provides users with a semantic search engine for data related to import and export control. In order to perform semantic analysis, the data are stored in the data lake in raw form. The BERT model is then used to convert the stored raw data into an embedding vector. Prior to text embedding, the BERT model receives text from the export control field and performs pre-training, followed by fine-tuning on the model's parameters using the dataset, which contains evaluations of semantic similarities between word pairs.

The BERT model encodes the search query into an embedding vector when it is entered. The similarity values between the vector and the cosine values of the existing data are then calculated, and the search results are returned in the order of highest similarity score. The degree of similarity between vectors is denoted by cosine values of the angle between two vectors. The cosine value approaches 1 as the similarity between the two data sets increases.

4. Conclusion

This paper conducted a conceptual analysis of an integrated export control database and its semantic search engine. The main idea behind creating an integrated database is to collect all enterprise data in a single central storage location known as a data warehouse and perform data analysis for each export control task using data marts. Furthermore, the semantic search engine will be implemented on a data lake system that stores unstructured raw data and converts it into the embedding vectors using a cutting-edge word representation model knows as BERT.

ACKNOWLEDGMENTS

This work was supported by the Nuclear Safety Research Program through the Korea Foundation Of Nuclear Safety (KoFONS) using the financial resource granted by the Nuclear Safety and Security Commission (NSSC) of the Republic of Korea. (No. 2106048)

REFERENCES

[1] The National Assembly of the Republic of Korea, Nuclear Safety Act, No. 16575, 2019.
[2] R. Kimball, M. Ross, The Data Warehouse Toolkit: the complete guide to dimensional modeling, John Wiley & Sons, 2011.
[3] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805, 2018.