Transactions of the Korean Nuclear Society Autumn Meeting
Gyeongju, Korea, October 29-30, 2015

The Role of Text Mining in Export Control

Jae-woong Tae, Choul-woong Son, Dong-hoon Shin*
Korea Institute of Nuclear Nonproliferation and Control., Yusungdae-ro 1534,Yusung-gu, Daejeon, Korea, 305-348
**Corresponding author: nucleo@kinac.re.kr*

1. Introduction

Nuclear items listed on Nuclear Suppliers Group (NSG) Guideline Part I are strategic items. It means that export licenses are required for transfers of Part I items. Moreover, Government Assurances are required. This leads to complicated administrative procedures and delays dates of export.

Exporters have to classify their items to strategic items and non-strategic items when they export nuclear items. They are punished if they export strategic items without export licenses. Korean government provides classification services to exporters.

It is simple to copy technology such as documents and drawings. Moreover, it is also easy that new technology derived from the existing technology. The diversity of technology makes classification difficult because the boundary between strategic and non-strategic technology is unclear and ambiguous.

Reviewers should consider previous classification cases enough. However, the increase of the classification cases prevent consistent classifications. This made another innovative and effective approaches necessary. IXCRS (Intelligent Export Control Review System) is proposed to coincide with demands. IXCRS consists of and expert system, a semantic searching system, a full text retrieval system, and image retrieval system and a document retrieval system. It is the aim of the present paper to observe the document retrieval system based on text mining and to discuss how to utilize the system.

2. Related Work

There are two kinds of approaches to classification, rule-based classification and case-based classification. Rule-based classification is a method that reviewers classify items according to a classification manual. Case-based classification is a method that reviewers consider previous classification cases when they perform classification.

Rule-based classification has been a major classification method. However, more attention is given to the case-based classification as reviewers are replaced and the number of classification increases.

The case-based approach requires long times and huge efforts. It also requires effective tools to deal with a lot of cases. Data mining is a technique to extract significant information from structured data. It is used in various fields including export control. However, review data and documents are text data which are unstructured data. In this case, text mining technique is applicable to this problem.

Text mining combines natural language processing and data mining. Text mining technique transform unstructured text data into structured data. Its application are document retrieval, document classification, document clustering, and information extraction.

The main approach in the field of text mining is a vector space model. Salton. G et al suggested term Frequency Inverse Document Frequency (TF-IDF) as keyword weighting method [1]. It depends on their term frequencies and document frequencies. The term frequency is the number of times a keyword occurs in a document. The document frequency is the number of documents in which the keyword occurs. The score for each keyword is higher when the term frequency is higher and the document frequency is lower. TF-IDF scores of documents form a vector space and the distance between vectors represents the distance between documents.

Text mining is also used in the real world. Ahn et al applied a text mining technique to LawnB which is an information system related to law and regulation [2]. This system shows that text mining technique is applicable to common systems and emphasized that a search result of a keyword-based retrieval system contains excessively many documents without considering its importance.

Go et al developed a patent retrieval system using text mining [3]. In this system, TF-IDF weighting and Euclidean distance measure were used.

3. Document Retrieval System

A document retrieval system based on text mining technique has been developed. It is embedded in IXCRS and serves to find similar documents to a new document.

The document retrieval system provides a list of documents in ascending order of document similarity. TF-IDF weighting and Cosine Similarity measure are applied to this system. Users input not a keyword but a document.

The system transform the format of a document such as PDF, PPT, DOC and HWP to TXT. Natural Language Processor create the list of keywords and its TF-IDF score which can be represented as a vector of TF-IDF score. Cosine Similarity is a method to measure

an angle between two document vectors. A small angle means two documents are similar.

The system has keyword lists and scores in a database concerning to NPP systems. It recommends three classes among 24 classes related to the nuclear power plant system matching keywords in a document and a database. Users can compare a new document to documents in selected classes or to all documents.

Fig. 1. A result of a class recommendation and a document retrieval

The System supports to input up to five documents together. It also provides similarities between input documents. However, Class Recommendation is omitted in this case. The greatest number of input documents is set up to prevent an overload. If necessary, it can be adjusted.

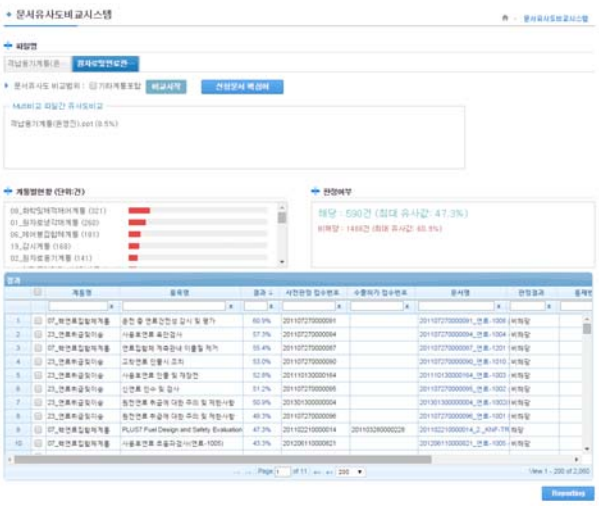

Fig. 2. A result of documents retrieval

A user can create reports using the system with checking similar documents. A report contain information about the retrieval result and a user can input opinion on the report.

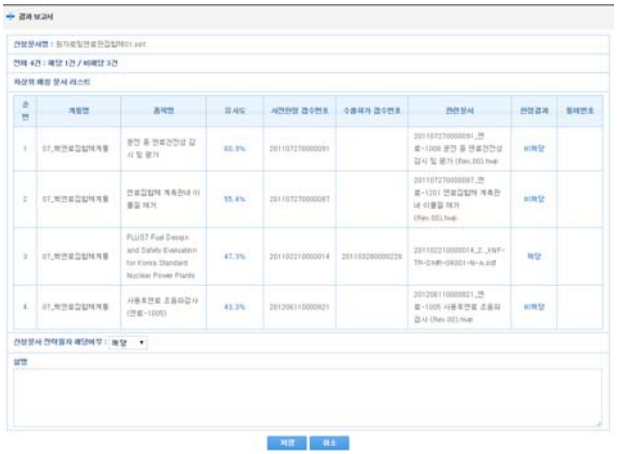

Fig. 3. An example of a report

4. Applications for Classifications

The document retrieval system can be used to utilize case-based classification. It explains a similarity between documents quantitatively and objectively. It handles many documents together and reduces the time required for searching documents dramatically. It also considers contents unrelated to the subject of a document. Without this system, reviewers should rely on their experiences and the subjects of documents.

However, the system has some limitation. Firstly, it cannot treat images. Documents consists of images and texts and images are important review factors. Especially, drawings including numerical data is strategic technology sometimes.

Secondly, the system cannot compare documents with different languages. English and Korean are used in most documents. English keywords and Korean keywords are treated as different keywords even if they have the same meaning. It will confuse exporters if a document and its translated documents have different classification results.

Thirdly, document similarity cannot represent inclusion relationships. The sizes of documents are different each other. Suppose that a small document are included in a large document such as Preliminary Safety Analysis Report (PSAR) which has contents related to most NPP Systems. Their similarity may become lower as the size of large documents increases. It means that it is hard to retrieve a very large document even if it contains similar contents.

For these reasons, reviewers should not rely on the system entirely and reviewers' opinion should have priority over the system. The system should play a role to remove loopholes as it sends signs to reviewers so that they pay special attention to classification. The number of strategic technology whose similarity is over

a threshold may be a useful index to measure possibility that a document is strategic technology.

Some data mining techniques such as artificial neural network and k-nearest neighbor (kNN) classification is also applicable to the system. They predicts the related system of a document and whether a document is strategic technology or not. It may be not perfectly precise but reviewers can refer the analysis results as a kind of indexes.

5. Conclusions

This study has demonstrated how text mining technique can be applied to export control. The document retrieval system supports reviewers to treat previous classification cases effectively. Especially, it is highly probable that similarity data will contribute to specify classification criterion.

However, an analysis of the system showed a number of problems that remain to be explored such as a multi-language problem and an inclusion relationship problem. Further research should be directed to solve problems and to apply more data mining techniques so that the system should be used as one of useful tools for export control.

REFERENCES

[1] Salton, G. and C. Buckley, "Term-weighting approaches in automatic text retrieval." Information Processing and Management: an International Journal 24(5): 513-523,1988

[2] 안태성, 서형국, 이경일, 텍스트마이닝 기반 고정밀 검색시스템, 정보처리학회지 제 11 권 제 2 호, 2004

[3] Gwang-su Go, Won-Kyo Jung, Young-Geun Shin, Sang-Sung Park and Dong-Sik Jang, “A Study on Development of Patent Information Retrieval Using Textmining”, Journal of the Korea Academia-Industrial cooperation Society Vol. 12, No.8 pp. 3677-3688, 2011