A Conceptual Design of the Information Analysis System for Searching Nuclear Fuel Cycle Related R&D Projects

Sung-ho Yoon and Dong-hoon Shin*
Korea Institute of Nuclear Nonproliferation and Control, 1534, Yuseong-daero, Yuseong-gu, Daejeon, Republic of Korea
*nucleo@kinac.re.kr

1. Introduction

Korea has concluded an additional protocol with the IAEA in 2004. In a part of additional protocol, information of nuclear fuel cycle related R&D projects supported by governmental funding should be reported to the IAEA. However, reports for some projects have been missed and IAEA occasionally found those missing by their information analysis system. This situation can cause distrust of the national transparence for nuclear nonproliferation. Therefore, we developed the system construction plan which can reduce some problems such as the report missing. This paper presents the result of the concept design as the first step of System development such as overall system structure, data collection and classification direction, post processing, and so on.

2. Related researches

There are lots of document categorization systems using computing based algorithm including machine learning. In the past, the majority of systems use the term-based classification method due to simple and powerful performance. But in recent, increasing computing power and processed big-data allow to develop real-time based adaptive method [1]. Our conceptual design is composed of previous methods and some of specialized functions to nuclear part.

3. Conceptual design

3.1 Overall system structure

The proposed system consists of four steps as shown in Fig. 1 including data collection, pre-processing, classification, and post-processing.

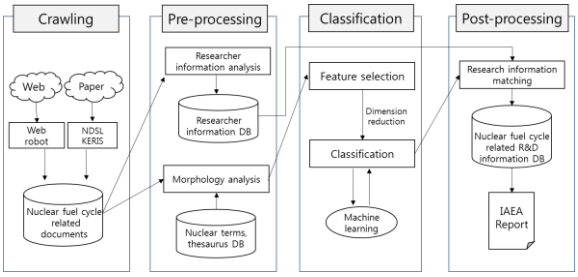

Fig. 1. Overall system diagram.

3.2 Data collection

At this stage, public research data (papers, project reports, technical reports, etc.) are collected from the open source such as website. In order to selectively collect data for nuclear fuel cycle, we use web-robot provided by search engines such as Naver or Google based on a pre-configured URL and a set of search keywords to automatically collect public information. For most papers that are not provided or published on the web, we will develop meta-search engine indirectly connected to the NDSL, DBPia, and so on.

3.3 Pre-processing

Most of the collected information will be in the form of a paper. Since the paper has a standardized format, at this stage, meaningful information such as title, author information (name, affiliation, e-mail address, etc.) and journal name can be acquired. In addition, only the title or abstract part containing the main contents of the paper can be set to the domain.

In the following, tags and abbreviations are removed and nouns necessary for feature selection

are extracted through morphological analysis.

3.4 Classification

Automatic document classification is roughly divided into two processes. The first is a feature selection process for converting initial data composed of a set of words to input data for machine learning. The second is a process for adjusting document classifier to target document group by conducting machine learning. Therefore, feature selection and machine learning algorithms are important for improving system performance.

3.4.1 Feature selection. If all the words are used as feature, it is difficult to conduct machine learning due to limited computing resource. Therefore, it is important to select the appropriate words that can reflect the characteristics of the document. There are several commonly used feature selection method such as Boolean weighting, TF/IDF, Information gain, Mutual information, etc.

3.4.2 Machine learning algorithm. There are several machine learning algorithms such as decision tree, artificial neural network, Bayesian network, genetic algorithm, etc. For this system, it is important to determine whether it is object or not. SVM (Support Vector Machines) [2] is an algorithm to select the most optimal hyperplane in the two-dimensional data classification problem as the decision boundary, therefore it seems that SVM meets the system purpose to distinguish between documents related to nuclear fuel cycle research and those that are not.

3.5 Post-processing

At this stage, the documents related to the classified nuclear fuel cycle are matched with the information of the researchers acquired in the pre-processing step and are automatically generated as a database and IAEA report.

4. Conclusion

In this study, a conceptual design was developed to construct a system that automatically classifies documents with a specific topic, nuclear fuel cycle research, from unspecific public document data.

There are several things to consider when constructing an actual system. First, the limitation of web robots for crawling public data. Recently, web robots are often blocked on internet search sites. Second, the difficulty of collecting and analyzing unstructured documents on the web. Because web documents are not formatted, character recognition and domain setting are difficult. Third, the optimization of machine learning algorithm. To customize algorithm to system a lot of training set are needed.

ACKNOWLEDGEMENT

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. 1803021)

REFERENCES

[1] W.H. Lee, S.J. Chung, and D.U. An, “Harmful Document Classification Using the Harmful Word Filtering and SVM”, Journal of Korea Information Processing Society, 1(16), 85-92 (2009).

[2] T. Joachims, “Text categorization with support vector machines: learning with many relevant features”, Proceedings of ECML-98, 10th European Conference on Machine Learning, 137-142 (1998).