Nuclear Exports Control System Using Semi-Automatic Keyword Extraction

Uihyun Kim, Hyunji Kim, Mun Yi, and Donghoon Shin

***Abstract*—Because the domain of nuclear power is highly specialized and complex, human experts have been utilized to manually evaluate all the documents submitted for export permission, causing the evaluation process to be slow and costly. Toward alleviating the problem of relying on laborious and costly human experts, the present research examines alternative approaches of text categorization, which is a key component of the case-based reasoning system proposed for the retrieval of documents only in the classes where a new export request case is related. Specifically, we examined three text categorization approaches: 1) manual approach involving a field expert, 2) automatic approach utilizing the TF-IDF scheme, and 3) semi-automatic approach involving both student experts and the TF-IDF scheme. Among the three methods, semi-automatic approach is the most efficient and effective in extracting keywords, demonstrating that the combination of machine and human is a promising solution that can effectively overcome the issues of expertise scarcity, time, cost, and accuracy simultaneously.**

***Index Terms*—Nuclear exports control system, case-based reasoning, text categorization, keyword extraction.**

I. Introduction

South Korea is very active in researching on nuclear power involving a variety of advanced reactors, fuel production and waste handling technologies, and power plant materials, seeking to export its nuclear technology, with a goal of exporting 80 nuclear reactors by 2030. To export a nuclear material, it is mandatory to obtain permission through a formal evaluation of whether the material is a strategic material because a strategic material is an essential material in a war even though it can be used for general industries as well. Today, the Korean government agency and other authorities have a great burden because the number of the evaluations of strategic materials is increasing rapidly with the growth of nuclear exports as shown in Fig. 1.

Up to now, however, there has not been any computerized system built to automate or aid the strategic material evaluation decisions. The current method, which is neither time-efficient nor cost-effective, requires field experts to manually analyze a large amount of domain specific documents to make the export permission decision. However, the size of the documents is too large to retrieve, usually leading to spending much time on examining irrelevant documents.

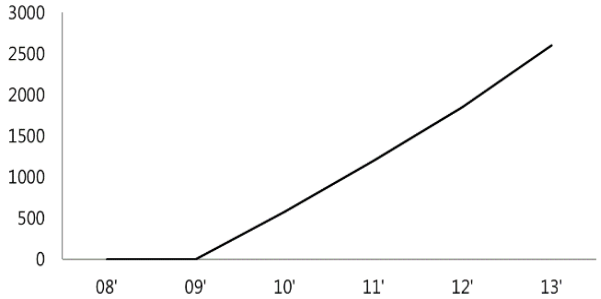

Fig. 1. The number of strategic materials evaluations in South Korea.

Within the context of building a case-based reasoning (CBR) system that facilitates the strategic material evaluation processes and decisions, we examine in this paper alternative approaches of text categorization. Text categorization is about the task of assigning electronic documents to one or more predefined categories, based on its contents. By categorizing nuclear technologies into a predefined set of classes depending on their properties, the CBR system can retrieve documents only in the classes where a new export request case is related, resulting in reduction of retrieval or indexing load.

For the proposed CBR system, we take a knowledge engineering approach in which the expert knowledge is directly encoded into the system in the form of production rules. It is the most capable approach, but it causes the knowledge acquisition bottleneck to occur when a small number of highly skilled experts should encode a large amount of knowledge-based rules [1]. Thus, this system examines the possibility of overcoming the bottleneck by comparing three keyword extraction techniques. Firstly, keywords are extracted automatically by TF-IDF, the most popular algorithm. Secondly, keywords are extracted semi-automatically, which the TF-IDF results are adjusted by student experts (who have no field experiences but studies nuclear engineering). Lastly, keywords are extracted manually by a field expert without any machine support. Once the system knows the keywords of classes, it can easily categorize a new case document by comparing a document and the keywords of each class.

Our study findings have direct implications for the building of CBR systems in general and strategic material evaluation systems in particular. The automated approach is the most cost-effective, time-efficient approach. The manual approach is the most labor-intensive, but traditional and reliable approach. The semi-automatic approach seeks to combine the strengths of both, without requiring the

Manuscript received March 3, 2014; revised April 15, 2014.

Uihyun Kim is with Tibero, Republic of Korea.

Hyunji Kim and Mun Yi are with the Department of Knowledge Service Engineering at KAIST, Republic of Korea (e-mail: munyi@kaist.ac.kr).

Donghoon Shin is with Korea Institute of Nuclear Nonproliferation and Control (KINAC), Republic of Korea.

involvement of field expertise, which is more costly and hard to obtain in a short period of time. The findings have the potential to solve the knowledge acquisition bottleneck issue while improving the performance of a CBR system.

II. Research Background

A. Text Categorization

Text categorization is one of the essential techniques in text mining, which is increasingly important research area as a plenty of text resources has been growing rapidly through the Internet. In recent years, it has been widely used in many applications such as text indexing [2], text sorting [3], text filtering [4], and cataloging web resources [5]. There are two main approaches to text categorization according to how to build classifiers [1]. Knowledge engineering approach manually builds classifiers from experts or expert reports. On the other hand, machine learning approach automatically builds classifiers by learning from a set of training data.

In the process of identifying a correct category for each document of nuclear material, the knowledge engineering approach is more suitable because of limited training data with security problems and necessity of an expert to focus on complex contents.

B. Knowledge Engineering Approach

The knowledge engineering approach is focused around manual development of categorization rules. A domain expert defines a set of rules for a document to be labeled with a given category. For example, Hayes developed CONSTRUE system to classify the Reuters [6]. In this approach, each rule is in the form 'IF Condition THEN Conclusion', where 'Conclusion' represents the appurtenance degrees to all predefined categories with logical structure. A typical rule in the CONSTURE is as follows:

IF <DNF (disjunction of conjunctive clauses) formula> **THEN** <category>

C. Knowledge Acquisition

Since the development of artificial intelligence and expert system, knowledge acquisition has been on the spotlight to capture expert knowledge. However, knowledge acquisition has still challenges to overcome the main problem called knowledge acquisition bottleneck. Furthermore, since the knowledge acquisition effort is highly resource intensive, the knowledge acquisition with insufficient data volumes needs to extract knowledge as rules directly form domain experts [7]. Thus, this paper proposes using keyword extraction to overcome problems from both the knowledge acquisition bottleneck and insufficient case data.

D. Keyword Extraction

Extracting keywords is to identify a set of terms that are the most relevant to the document, and it can be done by a human indexer or a machine. Nowadays, the performance of machine has been improved thanks to text mining and in particular TF-IDF, which is widely used to extract keywords that appeared frequently in a document. However, in general, manual indexing has not been substituted by automatic indexing with its high quality and excellent precision [8].

III. Experiment

A. Data Set

The experiments were conducted on two sets of documents from KINAC (Korea Institute of Nuclear Nonproliferation and Control) — training data and test data.

A training data consisted of a collection of 134 types of nuclear system manuals described in text. To set knowledge bases about 134 classes, for each type, three sets of keywords were extracted from the documents through three methods of the keyword extraction.

A test data consisted of a collection of 46 nuclear export documents also described in text. This data was used to see how well the text categorization works by finding an appropriate category for each nuclear export request proposal.

B. Preprocessing

Before extracting the keywords, preprocessing was conducted using Korean morphological analyzer [9] to transform words in a raw unstructured data source into a form of words that is available to analyze.

POS (Part-Of-Speech) tags divide words into categories based on the role they play in the sentence in which they appear—article, noun, verb, adjective, preposition, number, and proper noun as shown in Fig. 2. In this paper, only noun (NN) and proper noun (NNP) are considered because keywords are mainly composed of them. After that, stemming which is a technique to reduce words to their grammatical roots was conducted so that the similar words were represented with a root term.

A class of sprinkler of reactor building is
only active class of heat removal of reactor building

Reactor (NN), Building (NN), Sprinkler (NN)
Class (NN), Active (NNP), Heat (NN), Removal (NN)

Fig. 2. An example of POS tags.

C. Knowledge-Based Rule

When new document categorization is carried out by counting the number of keyword matches, it is assigned to the category that has the most number of keyword matches according to the IF-THEN rules made in reference to the categorization systems [6]. Note that, 'S_i' is the number of keyword matches between a category 'C_i'and a new document. In Fig. 3, suppose that a new document contains all 5 keywords of C_2 with the highest scores, then C_2 is the most appropriate category of the document. If multiple categories are ranked at the top, the document is assigned based on the comparison of the similarity of their titles.

D. Keyword Extraction

Keyword extraction is an important technique to find more relevant categories. In this paper, three keyword extraction approaches were used to achieve high-performance in text categorization—manual, automatic, and semi-automatic indexing. We limited the number of key-words extracted with three keyword extraction methods to five, following Turney [10], who limited the number of keywords extracted with his GenEx system to five. In addition, many academic

journals ask their authors to provide a list of about five keywords.

1) Manual keyword extraction, field expert

It takes a long time to become an expert in nuclear export evaluation domain. In this study, a very experienced field expert, who is currently working in KINAC for the evaluation of the export of nuclear systems, was asked to manually assign keywords to each class. He manually read whole documents and assigned five keywords to each category, taking about 9 hours for the whole task.

2) Automatic keyword extraction, TF-IDF

TF-IDF extracts keywords that appear frequently in a text. TF-IDF is composed of two components — TF (term frequency) and IDF (inverse document frequency). TF value means that more frequent words in a document are more important than less frequent words, and IDF value represents rarity across the whole document collection [11]. For example, although a word 'hot' appears frequently in a document, this word would be less important if the word also appears frequently across other documents.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (1)$$

where $n_{i,j}$ is the number of occurrence of the considered term in document d_j, and $\sum_k n_{k,j}$ is the number of occurrences of all term in document d_j.

$$idf_i = \log \frac{|D|}{|\{d_j: t_i \in d_j\}|} \quad (2)$$

where $|D|$ is total number of documents in the corpus, and $|\{d_j: t_i \in d_j\}|$ is the number of documents where the term t_i appears.

$$tfidf_{i,j} = tf_{i,j} \times idf_i \quad (3)$$

IF the most number of keyword matches
THEN $\{(C_1, S_1), \ldots, (C_{134}, S_{134})\}$

Categorization=$\{(C_1, S_1), \ldots, (C_{134}, S_{134})\}$

Category: C_2. Reactor Coolant System
Keywords: coolant, reactor, vessel, coolant pump, full core

Example: If Categorization=$\{(C_1, 3), (C_2, 5), (C_3, 2), \ldots, (C_2, 4), (C_{134}, 1)\}$

A new document belongs to the category C_2(Reactor Coolant System) because of the highest weight.

Fig. 3. Example of knowledge-based rule.

3) Semi-automatic keyword extraction, student experts with TF-IDF

This is an approach combined the advantages of automatic and manual approach. This is done by two steps of keyword extraction. At first, top-50 candidate keywords were extracted through TF-IDF algorithm. Next, student experts, who majored in nuclear systems at least over three years in college, removed meaningless terms and then selected final keywords by comparing candidate keywords that were ranked within the top-50 suggested by TF-IDF. Finally, for a consistency, they went over the results of keyword extraction generated by themselves and finalized the results. The whole process took 3 hours consisting of one and a half hour for extracting keywords, a half hour for break, and another hour for discussing their results and resolving the differences. By doing so, the machine approach (TF-IDF) accelerated the keyword extraction process and the human approach was expected to improve the accuracy of the term selection.

E. Overall Process

In the beginning, five keywords for each of the 134 classes were extracted from a training data using manual, automatic, and semi-automatic keyword extraction approaches described above, and then the keyword sets and the categorization knowledge-based rules comprised the knowledge base of the CBR system.

For the evaluation of the text categorization, once a new document in the test data came in, the system identified the most appropriate category where the most number of keywords in the category was also found in the new document according to the knowledge-based rule. More specifically, after the new document was compared against each of the 134 categories, the category that shares the most number of keywords with the new document was selected. Then, the new document was assigned to the best-matched category. When multiple categories were ranked together as the top-matching category, the new document was assigned to those multiple categories. After assigning all of the 46 test documents, the classification results were compared against the answers prepared by the field expert in advance. This evaluation process was identical for the three keyword-extraction approaches. The overall process of the experiment is graphically shown in Fig. 4.

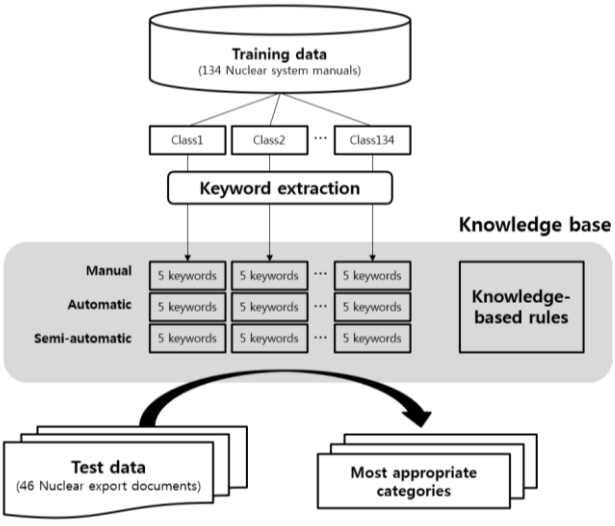

Fig. 4. Overall process of experiment.

IV. Evaluation Result

A. Evaluation Measures

To evaluate the performance of the three keyword extraction approaches, three measures, which is commonly used to evaluate ranking based document retrieval systems, were used—precision, recall, and F-measure [12].

Precision is the fraction of retrieved instances that are relevant while recall is the fraction of relevant instances that are retrieved. F-measure combines the two measures by multiplying them. For text categorization, given a classifier whose input is a document and whose output is a ranked list of categories assigned to that document, the recall and precision can be defined with the terms in Table I. The terms positive and negative refer to the prediction of classifier, and the terms true and false refer to whether that prediction corresponds to the external judgment. Thus, precision is the percentage of category assignments that were actually done correctly as in (6), while recall is the percentage of the total number of times that a particular category should have been assigned to a text and was in fact assigned as in (7). In this research, we calculated precision at 1 because we only search the most appropriate category for a document. We also calculated F-measure, which is the harmonic mean of precision and recall with evenly weighted as in (8).

TABLE I : MODEL FOR CLASSIFER ACCURACY

	Actual class	
Predicted class	tp (true positive) correct result	fp (false positive) unexpected result
	fn (false negative) missing result	tn (true negative) correct absence of result

$$\text{precision} = \frac{tp}{tp+fp} \tag{6}$$

$$\text{recall} = \frac{tp}{tp+fn} idf_i = \log \frac{|D|}{|\{d_j : t_i \in d_j\}|} \tag{7}$$

$$F = 2 \times \frac{precision \times recall}{precision + recall} \tag{8}$$

B. Evaluation Results

Table II summarizes the results of the three approaches evaluated by the three measures. Labeling data is used to extract keywords for each category, and test data is used to see how well test document is assigned correctly to category as real expert would do. A set of correct answer is provided by a very experienced KINAC expert who really does this works in real. Comparing keywords of categories assigned using labeling data and keywords of test document, test document will be assigned to a specific category that shares the highest number of common keywords. If the result of assignment is same with the correct answer set, it scores as defined in evaluation measure section. Since keywords extracted by three approaches are different, evaluation scores are also vary.

TABLE II: EVALUATION RESULTS

Measure	**Keyword extraction method**		
	Manual	***Automatic***	***Semi-automatic***
Precision at 1	0.434	0.370	**0.500**
Recall	0.426	0.362	**0.489**
F-measure	0.430	0.366	**0.495**

Overall, the semi-automatic approach achieved the average of 15% higher performance than the manual approach, which achieved the average of average 17% higher performance than the automatic approach, and semi-automatic approach achieved the average of 35% higher performance than the automatic approach. In sum, semi-automatic approach was the most effective keyword extraction method among the three examined, followed by the manual approach.

V. CONCLUSION

In this paper, we propose a novel information system in the domain of strategic material decision making, where it is difficult to rely on totally automated processes. The proposed system categorizes incoming documents by extracting keywords manually, automatically, or semi-automatically and finds relevant documents by applying knowledge-based rules. Among the three keyword extraction approaches, the semi-automatic method demonstrated its superiority compared to the other approaches. Solving the problem of knowledge acquisition bottleneck, the combination of machine and human provided a significant advantage in savings for expert labor time and effort while improving the accuracy of categorization by 35%. The semi-automatic approach has proved to be in average 15% better than the manual approach even when those student experts were much less knowledgeable about the domain than the field expert in the manual extraction condition. We think that semi-automatic approach delivers the best performance because this approach takes advantage of both machine and human, and combines them in a synergetic way. The machine component of TF-IDF screens out marginally relevant keywords, allowing student experts to focus on more selected keywords. Further student experts had the benefit of group discussion, cross checking each other's work through exchange of ideas and opinions. Thus, the machine approach implemented with TF-IDF contributed to reducing the time and consideration set while the human approach using student experts contributed to improving the accuracy of the term selection.

In order to obtain more verified and generalizable results, there is a need to increase the size of the data. Due to the security issue of the sensitive information, it was difficult to obtain a larger set of training data. In addition, because the accuracy of the Korean morphological analyzer used for the study was about 70% [13], the overall result of precision and recall was not as higher as other research that used English morphological analyzers, which have much higher accuracy rate. Notwithstanding these limitations, however, the study results clearly show that semi-automatic approach is a highly promising approach that can effectively overcome the issues of expertise scarcity, time, cost, and accuracy simultaneously.

ACKNOWLEDGMENT

Research reported in this paper was supported by Korea Institute of Nuclear Nonproliferation and Control (KINAC) and financially supported by Nuclear Safety and Security Commission (NSSC) as 2013 nuclear safety research project (1305014-0113-SB110).

REFERENCES

[1] R. Feldman and J. Sanger, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, CA: Cambridge University, 2006.

[2] G. Salton, A. Fox, and H. Wu, "Extended Boolean information retrieval," *Communications of the ACM*, vol. 26, no. 11, pp. 1022-1036, 1983.
[3] J. Hayes, E. Knecht, and J. Cellio, "A news story categorization system," in *Proc. the second conference on Applied Natural Language Processing*, pp. 9-17, 1988.
[4] H. Drucker, D. Wu, and N. Vapnik, "Support vector machines for spam categorization," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1048-1054, 1999.
[5] G. Attardi, A. Gulli, and F. Sebastiani, "Automatic Web page categorization by link and context analysis," in *Proc. THAI*, vol. 99, no. 99, pp. 105-119, 1999.
[6] J. Hayes and P. Weinstein, "Construe-Tis: A System for Content-Based Indexing of a Database of News Stories," *IAAI*, pp. 49-64, 1990.
[7] B. Buchanan and R. Smith, "Fundamentals of expert systems," Annual Review of Computer Science, vol. 3, no.7, pp. 23-58, 1988.
[8] A. Hulth, J. Karlgren, A. Jonsson, H. Boström, and L. Asker, "Automatic keyword extraction using domain knowledge," *Computational Linguistics and Intelligent Text Processing*, pp. 472-482, 2001.
[9] Intelligent Data Systems Laboratory in SNU. [Online]. Available: http://kkma.snu.ac.kr
[10] P. Turney, "Learning to extract keyphrases from text," Technical report, Na-tional Research Council, Institute for Information Technology, 1999.
[11] N. Aizawa, "An information-theoretic perspective of tf-idf measures," *Inf. Process. Manage.*, vol. 39, no. 1, 2003, pp. 45–65.
[12] Y. Yang, "An evaluation of statistical approaches to text categoriza-tion," *Information Retrieval*, vol. 1, no. 1-2, pp. 69-90, 1999
[13] W. Holsapple and D. Joshi, "Knowledge management: A threefold framework," *The Information Society*, vol. 18, pp. 47-64, 2002.

Uihyun Kim was born in the Republic of Korea in 1986. He received his master's degree in knowledge service engineering from KAIST, Republic of Korea, in 2014, and bachelor of electronic engineering from Hongik University, Republic of Korea, in 2011. His research interests include intelligent systems, HCI, and big data. He is working as a big data analyst at Tibero.

Hyunji Kim was born in the Republic of Korea in 1990. She received her Master's degree in knowledge service engineering from KAIST, Republic of Korea, in 2014, and bachelor of computer science and engineering from Seoul National University of Technology, Republic of Korea, in 2012. Her research interests include UX, HCI, and intelligent systems. Currently she is a doctoral student in the department of Knowledge Service Engineering at KAIST.

Mun Yi was born in the Republic of Korea in 1963. He earned his Ph.D. in information systems from University of Maryland, College Park. His current research interests include technology adoption and diffusion, IT training, knowledge engineering, and semantic Web. He is a professor in the department of Knowledge Service Engineering, KAIST, where he currently serves as the department chair. He is an associate editor of *International Journal of Human-Computer Studies* and a senior editor of *AIS Transactions on Human-Computer Interaction*.

Donghoon Shin was born in the Republic of Korea in 1976. He received his master's degree in medical physics from Catholic University and Ph. degree in nuclear engineering from Seoul National University, Republic of Korea, in 2007. His research interests include the data & text mining, artificial intelligence, image similarity, and applications for nuclear nonproliferation policy and implementation. He is working as a senior researcher in Korea Institute of Nuclear Nonproliferation And Control (KINAC).