Keyword Clustering for Comparing Documents in Different Languages

J. Tae and D. Shin

***Abstract*—The objective of this study was to complement natural language processing of a content-based retrieval system by applying keyword clustering. We focused on comparing documents in two languages. To evaluate the performance of this approach, we clustered keywords using the features of documents and performed document clustering using the results of keyword clustering. The purity and the entropy of document clustering revealed that keyword clustering resulted in improvements in the quality of document clustering and allowed us to measure similarities between documents in different languages.**

***Index Terms*—Keyword clustering, dictionary, document clustering, purity, entropy, export control.**

I. Introduction

Few countries have the capability for a full nuclear fuel cycle, and the great majority must import major items as part of the development of a civilian (or military) nuclear program. The nuclear power industry and other peaceful uses of nuclear technology require international transactions involving the materials, equipment, and technologies that could contribute to development of a nuclear weapon [1]. The same applies to munitions, missiles, and bio-chemical weapons. Such items, termed strategic items, can be used for the development, production, and applications of weapons of mass destruction (WMDs). The export of strategic items is subject to control according to the UN Security Council (UNSC) resolution 1540, multilateral export control regimes (MECRs), and a number of other treaties.

MECRs are international bodies that control strategic items; the four major MECRs are the Wassenaar Arrangement (WA), the Nuclear Suppliers Group (NSG), the Australian Group (AG), and the Missile Technology Control Regime (MTCR). The Nuclear Suppliers Group (NSG) was established in 1978 among seven suppliers of nuclear material: Canada, France, the Federal Republic of Germany, Japan, the Soviet Union, the United Kingdom, and the United States. It is an informal group that seeks to prevent the acquisition of nuclear weapons by countries other than those recognized as nuclear weapon states by the Nuclear Nonproliferation Treaty (NPT) [1]. As of 2014, the NSG had 48 member states.

The NSG published guidelines consisting of Parts I and II. Each guideline includes a list of controlled items. Some items in Part I are also termed trigger list items because the transfer of Part I items may trigger International Atomic Energy Agency (IAEA) safeguards. Moreover, it also requires physical protection. The government of the exporting country exchanges official letters with the government of importing country to assure safeguards and physical protection when trigger list items are exported abroad. Part II items are used in the field of non-nuclear industries as well as nuclear industries. Transfers of Part II items do not require the above process. In this respect, the transfer of Part I items is more complicated than that of Part II items.

A. Export Control of Nuclear Items in South Korea

The Nuclear Safety and Security Commission (NSSC) is the government agency responsible for export control of Part I items in the Republic of Korea. Each exporter who plans to transfer NSG Part I items must obtain licenses from the NSSC according to the Foreign Trade Act. If an exporter is not convinced that the items intended for transfer are not strategic items, they can request strategic item classifications from the NSSC.

The Korea Institute of Nuclear Nonproliferation and Control (KINAC) supports NSSC provision of technical information and expert opinion. It also supports the operation of the Nuclear Export Promotion System (NEPS), which is an online system to increase the efficiency and convenience of export control. Nuclear industry exporters apply for classification through the Nuclear Export Promotion System (NEPS). NSSC and KINAC then classify the items and issue export licenses via NEPS.

The export of a nuclear power plant was a turning point for export control in South Korea. On 27 December 2009, the United Arab Emirates selected a consortium led by Korea Electric Power Corporation (KEPCO) to design, build, and assist in the operation and maintenance of four 1,400-MWe nuclear power plants, i.e., the Advanced Power Reactor 1400 (APR1400) [2]. This made it difficult to implement timely export control because it involved many transfers of nuclear items, and the manpower of NSSC and KINAC was limited. Additionally, there was no absolute standard for strategic items. In the NSG guidelines, Part I items are defined as especially designed or prepared (EDP) equipment and components; however, the concept of EDP is ambiguous and subjective.

Nuclear items require a high level of specification and strict quality control because a nuclear power plant is a mission-critical system, and its safety is extremely important. Moreover, many of the components must withstand the high-pressure and high-temperature environment of the reactor coolant. Major components of the APR1400, including the reactor vessel, steam generator, and reactor coolant pump, are subject to control according to the NSG guidelines. However, this does not imply that all nuclear

Manuscript received November 20, 2014; revised February 26, 2015. This work was supported in part by Nuclear Safety and Security Commission (NSSC) and Korea Radiation Safety Foundation (KORSAFE).

The authors are with the Korea Institute of Nuclear nonproliferation and control (KINAC), 1534 Yuseong-daero, Yuseong-gu, Daejeon, 305-348, Republic of Korea (e-mail: ttjjww@postech.ac.kr, nucleo@kinac.re.kr).

items are EDP items. For example, a pressurizer is not subject to control, even though it is a critically important component of the APR1400 and is required to maintain the pressure.

NSG-participating governments have clarified EDP items through many discussions and meetings. However, this requires much time to clarify the meaning of EDP, and the NSG may require revision and the controlled item list only under the agreement of all participating states. The current guidelines do not provide NSG members with criteria that are sufficiently clear to identify strategic items. Therefore, each government requires its own policy to identify strategic items.

The necessity of individual sets of criteria to identify strategic items has been emphasized by the NSSC since the export of APR1400. The NSSC and KINAC have attempted to increase the consistency of export control implementation. As a result, the ability to retrieve similar documents and analyze ambiguous classification cases is regarded as a key factor. However, NEPS has limited retrieval performance and struggles to cope with the large number of components of the APR1400. To address this, the Intelligent Export Control Review System (IXCRS) has been proposed to analyze classification cases.

B. IXCRS

IXCRS is an expert system designed to support reviewers and government officers responsible for export controls. IXCRS, which has been actively developed since 2012, uses approaches including text mining, ontology, the Semantic Web, and image processing. IXCRS includes a document retrieval system that was developed based on term frequency–inverse document frequency (TF-IDF) and cosine similarity and that allows the user to find similar documents.

TF-IDF scores keywords depending 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 [3]. TF-IDF and cosine similarity can be used to describe the general similarity between documents. The distribution of keywords is another important factor describing documents; however, this approach is limited in some cases.

The documents contained in NEPS are written in several languages, although most are written in Korean or English. Korean and English keywords with the same meaning differ in the current system, and users cannot find similar documents in different languages using the current system.

Here, we consider the meaning of keywords to measure the similarity between documents in different languages. A Korean–English dictionary is used to establish links between Korean and English keywords. We approach this problem using keyword clustering.

II. Methods

In the field of computer vision, the "bag of words" model is commonly used to measure similarity between images [4]. A variety of image features, such as scale-invariant feature transform (SIFT) [5] and speeded-up robust features (SURF), have been reported [6]. These techniques can be used to represent the characteristics of images. SIFT detects local descriptors, where each descriptor is a 128-dimensional vector with scalar values in the range 0–255. A descriptor acts as a keyword in the field of computer vision. In contrast to text mining, it is not straightforward to say that two descriptors are the same because of the high dimensionality. It is common to cluster descriptors and to use this as a dictionary to overcome this problem. With this approach, two similar descriptors can be treated as the same using a dictionary. We adopted a similar approach here.

Korean–English dictionaries typically do not contain sufficient words for effective use in nuclear engineering. The names for specific items and technical terms are important keywords, but they are not commonly found in general-purpose dictionaries.

In this study, we assumed that keywords inherit features from the documents in which they occur. Keyword matching was performed using these features, and in this manner, a dictionary could be built. Dictionaries can be applied to various text-mining algorithms, such as classification, clustering, and retrieval. Document clustering was chosen because it can describe the similarity between documents in different languages, and can also identify whether they have similar properties.

Various metrics can be used to determine the quality of clustering, including purity and entropy. To evaluate the proposed method, we here performed document clustering and calculated the purity and entropy. Clustering and measurements of the purity and the entropy have been used to analyze the effectiveness of similarity measures, such as the Euclidean distance, Jaccard coefficient, and cosine similarity [7], [8]. Here, we used only the cosine similarity based on TF-IDF because this is the most general measurement of similarity, and we took this as the focal point of clustering of keywords and documents. We did not consider the performance of Natural Language Processing (NLP) for the same reason.

Purity describes the coherence of a cluster. A feature of a cluster can be defined as the feature of an element. For example, the class of a cluster can be defined as the class in which the number of elements is the largest. The purity of a cluster increases as the number of elements with the same class as the cluster increases.

Entropy also describes the coherence of a cluster. However, it differs from purity in that it considers the distribution of all classes in a cluster. Two clusters may have different entropies when a given feature has more than three values. Entropy also considers the distribution of the entire cluster. The entropies of two clustering results may differ even if they have the same purity. Lower entropy with the same purity means that the number of clusters is larger and the numbers of elements of the clusters are more unbiased. High purity and low entropy typically correspond to a good clustering result.

Let the number of elements of a cluster be n, let n_i be the number of elements of a cluster that belongs to the i^{th} class, and let m be the number of classes. The purity and entropy of the cluster are as follows:

$$\text{Purity}(\text{Cluster}) = \frac{\max(n_i)}{n} \quad (1)$$

$$\text{Entropy(Cluster)} = -\frac{1}{logm}\sum_{i=1}^{m}\frac{n_i}{n}log\frac{n_i}{n}$$

The entropy of clustering is obtained by summing the entropy of all clusters weighted by the size of each cluster. Let k be the number of clusters, h be the total number of elements, and h_i be the number of elements of the i^{th} cluster. Then, the purity and entropy of the clustering results are given by

$$Purity = \sum_{i=1}^{k}\frac{h_i}{h}\text{ Purity}(i^{th}\text{ Cluster})$$
$$Entropy = \sum_{i=1}^{k}\frac{h_i}{h}\text{ Entropy}(i^{th}\text{ Cluster}) \quad (2)$$

Lower purity may correspond to better performance in some cases. For example, consider a cluster consisting of a section in Korean and the corresponding section in English. This is an ideal case for clustering pairs of translated documents. Language is a major feature of the documents, and low purity corresponds to good performance for this feature. We define impurity as follows for convenience:

$$Impurity = \frac{1-purity}{1-\frac{1}{m}} \quad (3)$$

A similar concept may also be applied to the entropy of the above language feature. High entropy is usually similar in meaning to high impurity; however, entropy for the language feature was not calculated because it was redundant, as the language feature had only two values, i.e., 'Korean' or 'English,' and the distribution of the number of clusters was not considered in this study.

A. Keyword Clustering

If two keywords belong to the same field, they may be expected to appear in the same documents frequently. Here, we assume that two keywords will have similar meanings if they have similar document distributions. Keyword clustering is the process of collecting similar keywords under this assumption. It has been applied to topic detection, whereby Dutch Wikipedia articles comprising 758 documents were analyzed, and topics were identified via clustering [9].

Keyword extraction is necessary for keyword clustering and document clustering, together with a measure for similarity or distance between keywords. Frotoma's keyword extraction software package was used with (NLP). Frotoma is a Korean company that deals with ontology and the Semantic Web. This package analyzes a document lexically and produces keyword frequency data and a keyword list. It also provides functionality to remove some stop words based on frequency information; however, the experiments were implemented without removing stop words because we did not have sufficient data on the stop words.

We used two keyword-matching methods. First, we matched keywords to the closest keywords in another language. This is a one-way matching method, and all Korean keywords were translated to English, or all English keywords were translated to Korean. Second, we clustered keywords using features. A similar clustering algorithm to single-linkage clustering was used, which does not define the distance between clusters, but rather considers distances between elements only. We used the cosine similarity as a measure of the distance without any weighting. The keywords were separated into two groups: Korean and English.

As with single-linkage clustering, the above algorithm transforms each element into a cluster, then agglomerates clusters containing a given element (let this be element 'A') to another cluster containing the element that is closest to element 'A' in the other group. A keyword and the closest keyword in the other group always belong to the same cluster. Thus, there is no cluster with only one element unless one of the groups is empty.

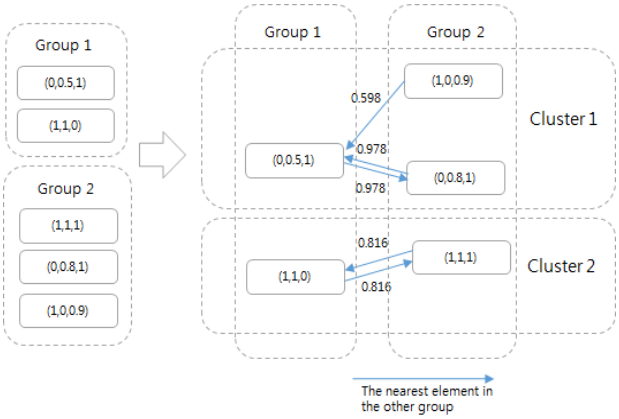

Fig. 1. An example of keyword clustering.

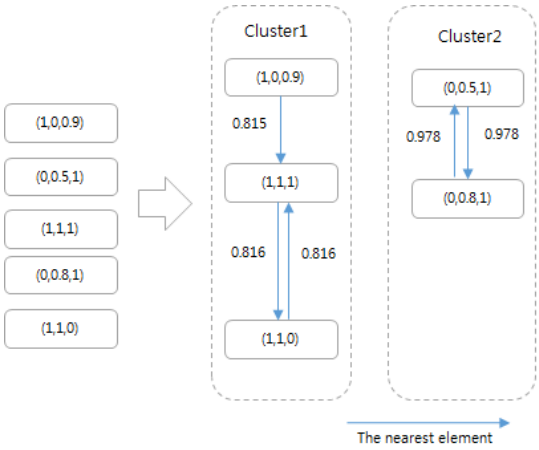

Fig. 2. An example of document clustering.

B. Document Clustering

Document clustering is a useful tool to evaluate the performance of features. A simple clustering technique is the single-linkage clustering algorithm, which defines the distance between two clusters and considers each element as a cluster. It iterates to find two clusters with the smallest separation and agglomerates them until n clusters are achieved. The TF-IDF and cosine similarity can serve as a measure of this distance because high similarity of two elements means that they are close.

We performed clustering using a modified single-linkage clustering algorithm, whereby each document was initially

considered a cluster, and a cluster containing the element 'A' was agglomerated with another cluster containing the element closest to 'A' [10]. This is similar to the above keyword matching clustering except for the groupings, as shown in Fig. 1 and Fig. 2.

With this document-clustering algorithm, a document and the closest document always belong to the same cluster. It follows that each cluster has at least two elements. However, it is not straightforward to predict the number of clusters, and the algorithm is not applicable if there are too many or too few clusters. We may expand clustering by agglomerating a cluster and clusters with the closest element, the next-closest element and so on. In this way, the algorithm becomes hierarchical clustering.

We will now consider how well document clustering separates documents into strategic items and nonstrategic items, and how similar the elements of a cluster are.

III. Experimental Data and Results

A. The Experimental Data

Two datasets of documents were used in the experiments. Dataset 'A' was made up of a large document containing over 8,000 pages and its translation. The documents in dataset 'A' were divided into 181 sections. Each pair of Korean and English sections had the same content. The language used is a major feature of documents in dataset 'A' (hereafter "the language feature"). The other dataset 'B' was a collection of 3,329 electronic documents in NEPS.

We constructed dictionaries for both datasets. Pairs consisting of a keyword and its representative keyword form the structure of a dictionary. The translation program substituted a keyword with its representative keyword prior to document clustering. If a keyword were not found in the dictionary, the clustering programs would not replace it with its representative; however, such cases did not occur because the dictionaries contained all keywords.

Dictionaries of dataset 'A' were constructed using the corresponding document frequency. The similarity measure was the cosine similarity with no weighting (such as TF-IDF). The section information to which a keyword belonged was a major feature used to match Korean and English keywords (hereafter referred to as "the section feature").

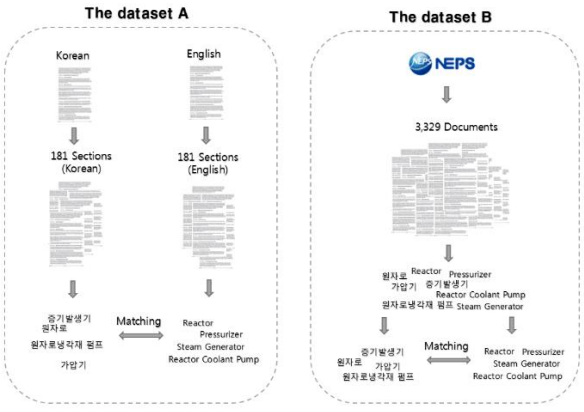

Fig. 3. The extraction of experimental data.

The number of unique keywords in dataset 'A' was 112,388. The number of keywords in each dictionary was 112,388 with clustering, i.e., 61,481 for Korean-to-English, and 55,625 for English-to-Korean translation. The number of representative keywords in each dictionary was 1,186 for clustering, 2,182 for Korean to English and 2,519 for English to Korean.

All documents in dataset 'B' were classified as to whether they were strategic items (this was not done for dataset 'A'). The documents were not distributed uniformly with reference to this feature; i.e., there were considerably more nonstrategic items than strategic items. Each of the documents in dataset B' were classified into 24 categories according to the related Nuclear Power Plant (NPP) systems. A nuclear power plant has several sub-systems, and these were used to describe the characteristics of documents. The number of NPP systems was large, so the 24-category structure was constructed by clustering sections of NPP system descriptions. We used this information as a keyword feature (hereafter "the NPP system feature"). However, the documents in dataset 'B' were not separated into Korean and English, and some documents were written in both languages. For this reason, all keywords were extracted from dataset 'B' and were divided into two groups: Korean and English. Keyword matching and keyword clustering were performed based on the NPP system feature. The numbers of keywords in each dictionary were 42,080 for clustering, 7,400 for Korean to English, and 36,941 for English to Korean. The number of representative keywords was 707 for clustering, 1,027 for Korean to English, and 1,234 for English to Korean.

B. First Experiment

The first experiment entailed clustering sections of dataset 'A' using three dictionaries and comparing the results with those obtained with no dictionary. The purity and entropy were calculated based on the chapters to which sections belonged. The impurity was computed using the language feature.

A cluster consisting of a section in Korean and the corresponding section in English represents an ideal case. The impurity for dataset 'A' relates to the language feature. Table I lists the impurity related to language. The impurity increased when the dictionaries were applied without significant deterioration of clustering performance; however, the quality improved when the clustered dictionary and the English–Korean dictionary were used. This shows that dictionaries contributed to comparing documents in different languages. The purity and entropy improved when the dictionary had more keywords. The clustered dictionary exhibited the best performance.

TABLE I: Results of Clustering Using Dataset 'A'.

Dictionary	Purity	Impurity	Entropy	No of clusters
Clustering	0.796	0.773	0.163	103
Kor->Eng	0.594	0.651	0.371	56
Eng->Kor	0.685	0.668	0.163	67
N/A	0.602	0.0608	0.275	54

We found better clustering performance when using the English–Korean dictionary than when using the Korean–English dictionary. This results from the

characteristics of the NLP. The NLP used in this study was better suited to processing Korean keywords because its developer was a Korean company.

The meanings of the matched or clustered keywords were not identical in a cluster. Moreover, many stop words were included in the dictionaries. In conclusion, the dictionaries were not adequate for use by humans.

C. Second Experiment

The second experiment was designed to examine how two dictionaries extracted from two groups of documents affect clustering. In this experiment, we performed document clustering of datasets 'A' and 'B' together. However, the Korean documents and the English documents of set 'A' were examined separately because it is possible that the pairs of sections in different languages form a cluster that excludes documents in set 'B' due to a characteristic of the clustering method. As a result, the total number of clustered documents was 3,510 in both cases.

It was not possible to build a dictionary from two datasets in this study because they did not have a common feature. We applied two dictionaries from each dataset sequentially. The dictionary from the dataset 'B' was given priority for translating keywords.

Documents of dataset 'A' tended to form clusters because they were from the same document, and they shared unique keywords of the source document. For this reason, the impurity was calculated considering how the documents from sets 'A' and 'B' were mixed. The impurity was significantly smaller compared with the first experiment. This is because the number of documents in the two datasets was so different. Theoretically, the maximum impurity was approximately 0.103134.

During classification of strategic items, the related NPP system and the conformance to strategic items were regarded as the most important factors. The purity and entropy were calculated based on these features after eliminating documents from dataset 'A' because they were not classified into both features.

There was no significant difference between the results of the two experimental sets, as shown by the data listed in Tables II and III. Applying the dictionaries improved the performance of clustering considering features related to NPP systems, as well as the conformance to strategic items. A clustered dictionary reflects major features related to NPP systems and strategic items better than does one-side keyword matching; however, the impurity was worse.

TABLE II: RESULTS OF CLUSTERING THE ENGLISH DOCUMENTS OF DATASET 'A' AND 'B'

Dictionary	Clustering	Kor->Eng	Eng->Kor	N/A
No of clusters	648	685	662	722
Impurity	0.0245	0.0330	0.0245	0.0291
Purity (NPP sys)	0.9303	0.8189	0.9204	0.8182
Entropy (NPP sys)	0.0501	0.1136	0.0568	0.1281
Purity (Strategic)	0.9330	0.9354	0.9336	0.7906
Entropy (Strategic)	0.1072	0.1105	0.1089	0.1127

TABLE III: RESULTS OF CLUSTERING THE KOREAN DOCUMENTS OF DATASET 'A' AND 'B'

Set	Korean documents of the dataset 'A' and the dataset 'B'	English documents of the dataset 'A' and the dataset 'B'
No of clusters	656	661
Impurity	0.0741	0.0838
Purity (NPP sys)	0.9285	0.9225
Entropy (NPP sys)	0.0523	0.0567
Purity (Strategic)	0.9213	0.9294
Entropy (Strategic)	0.1246	0.1145

A major obstacle to clustering with dataset 'B' was that the similarity between a Korean document and an English document was very low, even if they had similar content, because they had few common keywords. We found that keyword clustering improved this problem and resulted in better document-clustering performance.

The impurity did not improve sufficiently, however, with the exception of just one case, even though we used two dictionaries from datasets 'A' and 'B'. It follows that independently extracted dictionaries from two sets of documents do not function well to concatenate datasets in general. If we want to match documents of set 'A' to those of set 'B' effectively, a common feature should be identified, and keyword clustering should be performed with both datasets using this common feature. However, there was no readily identifiable common feature of the two datasets.

D. Third Dxperiment

The third experiment was similar to the second experiment; however, different dictionaries were applied. We attempted to improve on the impurities of the second experiment by applying different dictionaries. To build a dictionary, 5,324 keywords that occurred in both datasets were identified by comparing keywords of the two datasets, rather than finding a common feature. These formed a mediator between the two groups of keywords, as they contained both the section feature and the NPP system feature. These were representative keywords of a new dictionary, and all keywords were matched to them using the cosine similarity without weighting. The representative keyword of each dictionary was the keyword with the greatest similarity. Document clustering was performed using this dictionary.

TABLE IV: RESULTS OF CLUSTERING WITH THE THIRD EXPERIMENT

Dictionary	Clustering	Kor->Eng	Eng->Kor	N/A
No of clusters	655	702	669	731
Impurity	0.0034	0.0256	0.0245	0.0262
Purity (NPP sys)	0.9306	0.8414	0.9219	0.8201
Entropy (NPP sys)	0.0505	0.1138	0.0554	0.1272
Purity (Strategic)	0.9339	0.9342	0.9342	0.7906
Entropy (Strategic)	0.1091	0.1131	0.1081	0.1115

As a result, these dictionaries increased the impurity, while increasing purity and decreasing entropy, achieving superior results compared with clustering without dictionaries. However, the performance was inferior compared with dictionaries in other experiments, except with regard to impurity. Compared with other experiments, some similar keywords were not matched because the relationships between representative keywords were not considered. We also found that the dictionary contained far more representative keywords.

Another interesting observation was that the representative keywords were more refined; i.e., many typographical errors were eliminated. This occurred because the probability of the same typographical error occurring in both datasets was very small.

IV. Conclusion

We have described the application of keyword clustering to compare groups of documents. In the first experiment, we considered documents in Korean and English, and in the second and third experiments, we compared 181 sections of a large document and 3,329 documents of NEPS. Document clustering was carried out based on dictionaries constructed via keyword clustering. The purity, impurity, and entropy of the clusters were calculated. Dictionaries were constructed with the restrictive condition that a common feature of keywords was absent in some cases.

The results show that keyword clustering can help not only in comparing documents in different languages, but also in establishing links between groups of documents when an appropriate dictionary is applied. We therefore recommend applying dictionaries to improve the performance of major features of strategic item classification in the field of export control. It is also necessary to account for common features of documents to achieve effective keyword clustering.

References

[1] I. Anthony, C. Ahlstrm, and V. Fedchenko. (2007). Reforming nuclear export controls: The future of the nuclear suppliers group. Stockholm International Peace Research Institute. [Online]. Available: http://book.sipri.org/

[2] S. Kim and J. Keppler. (2013). The Barakah Nuclear Power Plants, The United Arab Emirate. presented at OECD NEA Workshop on Electricity Prices and Nuclear New Build, Paris. [Online]. Available: http://www.oecd-nea.org/

[3] G. Salton and C. Buckley. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management: an International Journal* [Online]. 24(5). pp. 513-523. Available: http://comminfo.rutgers.edu/~muresan/IR/Docs/Articles/ipmSalton1988.pdf

[4] C. Tsai. (2012). Bag-of-words representation in image annotation: A review. *ISRN Artificial Intelligence*. [Online]. pp. 1-12. Available: http://www.hindawi.com/journals/isrn/2012/376804/

[5] D. G. Lowe. (2012). Distinctive image feature form scale-invariant key. *ISRN Artificial Intelligence*. [Online]. pp. 1-12. Available: http://www.cs.berkeley.edu/~malik/cs294/lowe-ijcv04.pdf

[6] H. Bay, T. Tuytelaars, and L. V. Gool. (2006). SURF: Speeded up robust feature. *European Conference on Computer Vision*. ECCV. [Online]. Available: http://www.vision.ee.ethz.ch/~surf/eccv06.pdf

[7] A. Huang. (2008). Similarity measures for text document clustering. *New Zealand Computer Science Research Student Conference*, [Online]. pp. 49-56. Available: http://www.milanmirkovic.com/wp-content/uploads/2012/10/pg049_Similarity_Measures_for_Text_Document_Clustering.pdf

[8] Y. Zhao and G. Karypis. (2004). Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning*. [Online]. 55(3). Available: http://dl.acm.org/citation.cfm?id=990398

[9] C. Wartena and R. Brossee. (2008). Topic detection by clustering keywords. *5th International Workshop on Text-based Information Retrieval*. [Online]. Available: http://www.uniweimar.de/medien/webis/research/events/tir-08/tir08-web/

[10] J. C. Gower and G. J. S. Ross. (1969). Minimum spanning trees and single linkage cluster analysis. *Journal of the Royal Statistical Society*. [Online]. 18(1). pp. 54-64. Available: http://adessowiki.fee.unicamp.br/media/Attachments/main/MainPage/mstSingleLinkage1969.pdf

Jae-Woong Tae was born in Pusan, Republic of Korea in September 1982. He received the MS degree from the Department of Mathmatical Sciences, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea, in 2010. He is a researcher in Korea Institute of Nuclear Nonproliferation and Control (KINAC). His current research interests include data and text mining, computer vision.

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.D degree in nuclear engineering from Seoul National University, Republic of Korea, in 2007. He is working as a senior researcher in Korea Institute of nuclear nonproliferation and control. His research interests include the data and text mining, artificial intelligence, image similarity, and applications for nuclear nonproliferation policy and implementation.