An Introduction to the Nuclear Document Crawling System

Jae-woong Tae, Sung-ho Yoon, 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

Detailed technologies concerned about operation, construction, design of nuclear reactors tend to be strategic technologies which are defined in the NSG (Nuclear Suppliers Group) guidelines [1] and to be subject to strategic trade control. However, it does not mean that all nuclear technologies are subject to control.

The NSG guidelines state that controls on "technology" transfer do not apply to information "in the public domain" or to "basic scientific research". According to the guidelines, "basic scientific research" is an experimental or theoretical work undertaken principally to acquire new knowledge of the fundamental principles of phenomena and observable facts, not primarily directed towards a specific practical aim or objective. "Technology in the public domain" means “technology” or “software” that has been made available without restrictions upon its further dissemination.

It is a difficult problem to determine whether a document is in the public domain or it is a basic scientific research because its criteria are ambiguous and unclear. In this paper, we introduce an approach using documents on the web and a system to manage electronic documents on the web.

2. Method

There are a lot of electronic documents related to the nuclear field on the web. In the view of strategic trade control, there are two viewpoints. One is that they are not subject to strategic trade control because they are open to public. The other is that someone transferred a nuclear document to unspecified individuals illegally if the document has detailed and important contents.

There is no absolute answer. However, most strategic trade control experts agreed to the former. Many nuclear businesses or researches cannot be performed by individuals but by enterprises and they have tried to prevent the leaking of their important technology. Research papers are similar to the above. Researchers do not publish a research paper including technology which is valuable commercially. Most technical papers provide theoretical approaches to design nuclear items or its safety evaluations.

We can approach this problem by utilizing nuclear documents on the web. Comparing a document to documents on the web provides us a list of similar documents and similarities. Reviewers can use this information for classification. If they finds out a very similar document on the web, they may classify the technology into non-strategic technology.

However, it is difficult that reviewers collect and manage a lot of documents on the web by themselves because it takes long time to search useful documents and to download them. Therefore, it is necessary to develop a document crawling system to apply this approach practically.

3. Document Crawling System

The nuclear document crawling system has been developed to deal with electronic documents in the nuclear-related field on the web effectively. The system consists of an automatic document crawling module, a manual document crawling module, a document retrieval module.

The automatic crawling module performs google search using keywords and web sites in the database regularly. It adopted Google API (Application Programming Interface) [2] to search documents and web pages on the web. The system downloads documents of URLs (Uniform Resource Locators) provided by Google API every day. It is unnecessary to explore the entire web because it is enough to explore documents only in the related fields such as nuclear engineering and chemical engineering. Google API provides data collected by the crawler of Google. It is free for hundred queries. A user can adjust the amount of files to download for one query by the ten. As this value increase, the system download more data but more noisy documents are also collected. Its initial value is ten.

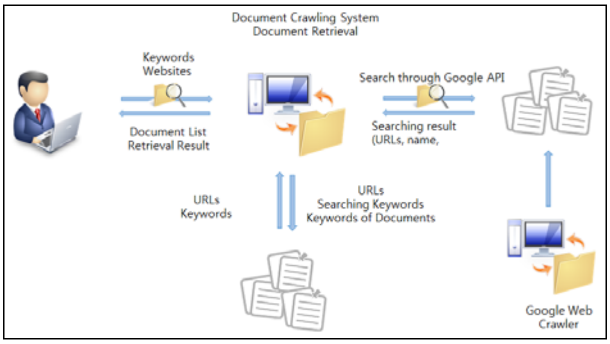
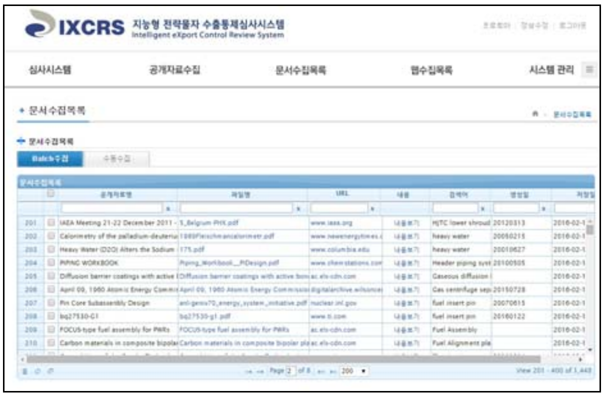


Fig 1. The flow of processes of the system

A user can browse through collected documents on the system. It provides its name, URL, date of creation, a search term and contents without formatting. Users can read its text but they must download the entire document again if they want to read its original document. It is an effective mean to save the storage of the system.

A User may find out an unimportant document in the list and can delete it. The system maintain its information practically because it is possible to collect deleted document again. The system just hides useless documents from the collected document list


IXCRS 지능형 전략물자 수출통제심사시스템
Intelligent eXport Control Review System
심사시스템
공개자료수집
문서수집목록
웹수집목록
시스템 관리

문서수집목록

문서수집목록

Batch수집 | 수동수집

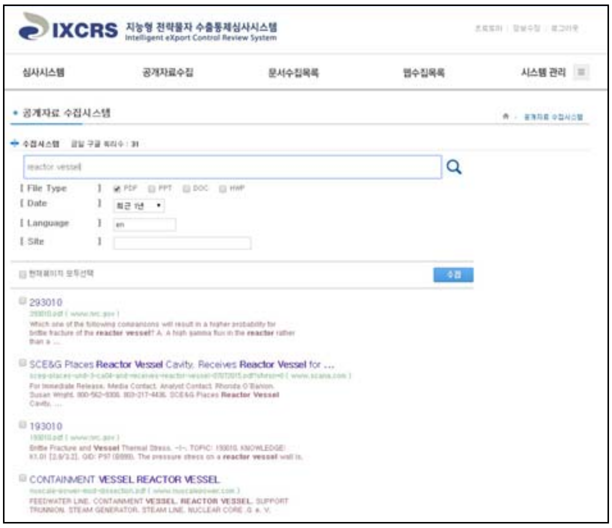
문서수집목록

		공개자료명	파일명	URL	내용	검색어	생성일	저장일
201	☐	IAEA Meeting 21-22 December 2011 -	5_Belgium PHX.pdf	www.iaea.org	내용보기	HJTC lower shroud	20120313	2016-02-1
202	☐	Calorimetry of the palladium-deuteriu	1989Fleischmanncalorimetr.pdf	www.newenergytimes.c	내용보기	heavy water	20050215	2016-02-1
203	☐	Heavy Water (D2O) Alters the Sodium	175.pdf	www.columbia.edu	내용보기	heavy water	20010627	2016-02-1
204	☐	PIPING WORKBOOK	Piping_Workbook__PDesign.pdf	www.chemstations.com	내용보기	Header piping syst	20100505	2016-02-1
205	☐	Diffusion barrier coatings with active	Diffusion barrier coatings with active bon	ac.els-cdn.com	내용보기	Gaseous diffusion		2016-02-1
206	☐	April 09, 1960 Atomic Energy Commi	April 09, 1960 Atomic Energy Commissio	digitalarchive.wilsoncei	내용보기	Gas centrifuge sep	20150728	2016-02-1
207	☐	Pin Core Subassembly Design	anl-genis70_energy_system_initiative.pdf	nuclear.inl.gov	내용보기	fuel insert pin	20070615	2016-02-1
208	☐	bq27530-G1	bq27530-g1.pdf	www.ti.com	내용보기	fuel insert pin	20160122	2016-02-1
209	☐	FOCUS-type fuel assembly for PWRs	FOCUS-type fuel assembly for PWRs	ac.els-cdn.com	내용보기	Fuel Assembly		2016-02-1
210	☐	Carbon materials in composite bipolar	Carbon materials in composite bipolar pla	ac.els-cdn.com	내용보기	Fuel Alignment pla		2016-02-1

Page 2 of 8 200 View 201 - 400 of 1,443

Fig 2. A list of documents collected automatically

The manual crawling module is designed for immediate document crawling. Automatic crawling is performed using small part of keywords in the database at the fixed time. A User needs to crawl documents immediately sometimes. The manual crawling module can use Google API up to twenty times since the automatic crawling module search less than eighty times in a day. Users can search electronic documents on the system like using Google. They can choose useful documents and download them together. Its result is more useful than documents collected automatically because users do not select noisy documents.


IXCRS 지능형 전략물자 수출통제심사시스템
Intelligent eXport Control Review System
심사시스템
공개자료수집
문서수집목록
웹수집목록
시스템 관리

공개자료 수집시스템

수집시스템 금일 구글 쿼리수 : 31

reactor vessel

[File Type] ☑ PDF ☐ PPT ☐ DOC ☐ HWP
[Date] 최근 1년
[Language] en
[Site]

☐ 현재페이지 모두선택 수집

☐ 293010
293010.pdf (www.nrc.gov)
Which one of the following comparisons will result in a higher probability for brittle fracture of the **reactor vessel**? A. A high gamma flux in the **reactor** rather than a ...

☐ SCE&G Places **Reactor Vessel** Cavity, Receives **Reactor Vessel** for ...
For Immediate Release. Media Contact. Analyst Contact. Rhonda O'Banion. Susan Wright. 800-562-9308. 803-217-4436. SCE&G Places **Reactor Vessel** Cavity. ...

☐ 193010
193010.pdf (www.nrc.gov)
Brittle Fracture and **Vessel** Thermal Stress. -1-. TOPIC: 193010. KNOWLEDGE: K1.01 [2.9/3.2]. QID: P97 (B999). The pressure stress on a **reactor vessel** wall is.

☐ CONTAINMENT **VESSEL REACTOR VESSEL**
FEEDWATER LINE. CONTAINMENT **VESSEL**. **REACTOR VESSEL**. SUPPORT TRUNNION. STEAM GENERATOR. STEAM LINE. NUCLEAR CORE. G. a. V.

Figure 3 A result of manual crawling

A document retrieval module is same to the document retrieval system of IXCRS (Intelligent Export Control Review System) [3]. It compare documents based on TF-IDF (Term Frequency-Inverse Document Frequency) [4] weighting and cosine similarity. However, it does not provide classification and it only suggests the list of documents in descending order of similarities. Classification into reactor system categories is useful only when all collected documents were classified but it is impossible.

Google API needs keywords for searching. Using Google API is equivalent to searching on the Google web page. The database of keywords is necessary to collect documents. Web sites data is helpful for precise search. There were much less noisy documents which is not related to the nuclear field when the API searches documents with website information together. The quality of search keywords and website information may determine the performance of the system and the quality of collected document database.

4. Applications in Strategic Trade Control

The developed system can be used in a classification process as aforementioned. It collects documents automatically based on search terms inputted by users. Reviewers can collect documents sufficiently through the manual crawling module.

Reviewers can upload documents to classify and compare it to collected documents. They can check the similarity between documents directly since the system provides a list of similar documents. They will classify a document into non-strategic technology if they find out similar documents. They will review a document in depth if they cannot find similar documents.

On the other hand, the system can be used to analyze the state of art trends related to nuclear nonproliferation. It has two document databases. One is for classification review and the other is for information analysis. There are also two sets of search terms which are managed separately. One set is composed of technical terms such as “a reactor system” and “a neutron detector” while the other consists of political terms such as “North Korea” and “UN sanction”.

Users can collect information related to nuclear industry, strategic trade control policies, nuclear activity of some countries, international concerns, opinions of specialized agencies and so on through this system. This information is not only used to establish non-proliferation policies but is also used for identifying proliferation risks and nuclear non-proliferation assessment in the export licensing.

The system provides reviewers with an efficient tool to manage the above information. Reviewers should perform web searches to collect important information regularly without it. It is inefficient since it takes long time to input many search terms on the web and it is possible to collect same information redundantly.

5. Conclusions

In this paper, we proposed an approach to determine whether a document is open to public or it is a basic scientific research and we developed the document crawling system to collect open documents on the web. We can take open documents into a review process in a new way. It supports to prevent reviewers from classifying an open document into a strategic technology. It is expected to improve reliability of classification results.

However, they are not refined sufficiently although a thousand of data has been collected. For example, there are many simple document that introduce a company which are useless in a classification review process. In future, filtering module should be developed to eliminate documents useless for reviewing and the keyword database should be complemented by adding new keywords and websites. In addition, some research papers are not for free or not available on the internet. The system cannot collect such papers. Such research papers should be collected manually to enhance the document database.

Acknowledgement

This work was supported by the Nuclear Safety Research Program through the Korea Foundation of Nuclear Safety (KOFONS), granted financial resource from the Nuclear Safety and Security Commission (NSSC), Republic of Korea (No.1305014)

REFERENCES

[1] Guidelines for Nuclear Transfers, INFCIRC/254/ Rev.12 / Part 1, IAEA, 2013

[2] S. Brin and L. Page, “The anatomy of a large-scale hyper-textual web search engine,” in Proceedings of the 7th International World Wide Web Conference, 1998.

[3] J. Tae, C. Son, D. Shin, The role of text mining in export control, Transactions of the Korean Nuclear Society Autumn Meeting, 2015

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