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 nonstrategic 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.

			1) 문서류시장	
44	REM.									
	381	82 81	171		840 8	4 7 4				
	3%	/85	[74]	Ac	978 3	ф.				
	215	172 01	服料做装订纸准备	85	V 4011 EX	비행세이니J,발로자성~0	내도), 풍성자감물?	Ladt		
A.	81 M									
		Actit	* 3050 8050 9750 1	19# 77#	#12 0 716 0 29	H.:				
0.71	8.01	4 08129 + 78	알보 자용추진							
		A9400	875							
	29			2480	1			- 10	11-58	chec
	1	05.80	(为会议从用服					5	3%	
	1	19.224	13.5					2	15	
	-		2월체학제어계를							
141	7 145	42.44	2124 424				•	1	15.	×
	145				+ 5150	8 0		1	05	×
718 01,31 13,27	145 261 452 411	M2 47			88	49 : 117건 (최대 음-) : 120건 (대대 류사))	42:23.7%)	\$	15.	×
718 01,31 13,27	14년 발원1 4일5 시개를 양파네	위교 사학 왕 (오라.건) 비학자에게 (고려고) 용 (175) 유럽자에게 등 (105)	0124 444	7	세명 비명말	: 117건 (최대 음) : 1992 (미대 음사))	4727 23.7%) 42.7%)	1		
· 개봉 01, 차 15, 값 05, 유	145 261 415 415	해고 4년 만(연박건) (전화6년 월 (15년) 월 (17일) 유럽44 해왕 (105) 제품왕	2124 #44	89 -	전달 비해당 4005 5+08	: 117전 (최대 유) : 129전 대의 유사값 : 428월 전의 유사값	42.7%) 42.7%) 843		2123	
· 제용 08.화 19.27 05.중	145 발전1 4일3 신시1 음	해고 4년 만(만박건) (123965개월 (154) 월 (173) 유감41개월 (105)	5324 #44		4003 0468	: 117전 (최대 유) : 129전 대의 유사값 : 428월 전의 유사값	422: 23.7%) 42.7%) 848	34	2122	
718 011, 27 011, 17 21	[45] 방안[신제] 알파네 () ()	(박고 4년 1 왕 (2위:건) (박지수) 위동 (254) 동 (175) 유급시 북동 (185) 가동왕	23124 #44	427%	400% 0440A	: 117전 (최대 유) : 129전 대의 유사값 : 428월 전의 유사값	427-23.7%) 42.7%) 8.4% 201407270000077	34	5235 10 2235	
· 제월 (81, 51 19, 22 (01, 61 2	145 발전1 박일리 신제1 일 (2) (2) (2)	(비교 사망 참 (2위:25) 비간(19) 등 (175) 유감시계를 (105) 가동원 (10.5) (10.5)	20274 404	a 42.7% 29.0%	4100% D+01A 100% D+01A 20+107270000077 20100420000741	: 117전 (최대 유) : 129전 대의 유사값 : 428월 전의 유사값	8 418 21 - 23, 7%,) 8 418 21 - 27 7000077, 21 0040000741	34 34 4710 KOSD 238	0122 a xN2 xN2	
718 011, 27 011, 17 21	145 201 425 425 441 344 344 344 344 344 344 344 344 344	(비교) 사외(참 (문제:건3) 비전340(개봉 (254) 용 (175) 유급자(개봉 (105) 가용문학 (1.25) 지용문학 (1.25) 지용문학 (1.25) 지용문학	C1112-14-1114	42.7% 29.0% 20.1%	415155 B+454 2010727000077 20100420000741 20100420000117	117전 (최대 유) 1320전 (제대 유사진) 수준비가 입수했요 ()*	(-2/- 23.7%) 42.7%) (84% (84% (84% (84%) (21/04/050007) (21/04/050007) (21/04/050007)	14 194 42710 4050 238	<25 <25 <25 <25 <25 <25 <25 <25 <25 <25	
· 제왕 08,51 15,22 05,6 2 3	145 방언1 학안 시비1 상자4 	(비교 시작 (전자6) 개를 (154) (175) 유감시개를 (155) (175) 유감시개를 (155) (15, 20시작을 (15, 20시작을 (15, 20시작을 (15, 20시작을 (15, 20시작을 (15, 20시작을 (15, 20시작을	(1974 404)) (1974 404)))	a 42.7% 29.0%	4100% D+01A 100% D+01A 20+107270000077 20100420000741	: 117전 (최대 유) : 129전 대의 유사값 : 428월 전의 유사값	84(8) 84	34 296 47010 4010 238 1010 238 3 Facel res	0122 + + + + + + + + + + + + +	
• 제월 08,51 19,22 05,6 1 2 3 4	145 방안1 식당 시세1 상자4 () () () () () () () () () () () () ()	응답고 시작[방 (241:23) 방 (1721) 방(1727) 유럽지 제품 (105) 가용 방 가용 방	C111274 4114	4275 2905 2975 2975	41555 5+654 11185 2010727000077 201002000017 2010020000017 2010020000017	117전 (최대 유) 1320전 (제대 유사진) 수준비가 입수했요 ()*	(23,7%) (24,7%) (21,1%	3 4 87310 4030 238 4030 239 5 Facel Int 6 71 8732	2222 422 422 422 422 422 422 422 424 42 42	
• 제월 08,58 19,27 05,8 24 3 4 5	145 발한1 박맛차 성자4 	(비교 시작 (전자6) 개를 (154) (175) 유감시개를 (155) (175) 유감시개를 (155) (15, 20시작을 (15, 20시작을 (15, 20시작을 (15, 20시작을 (15, 20시작을 (15, 20시작을 (15, 20시작을	C111274 4114	427% 290% 297% 297% 218%	40555 0+40 100000077 201002000011 201002000011 201001000011 20100110000011 20101110000011	1172 (3)(3) (3) 1723 (3)(3) (3)(3) 1723 (3)(3) (3)(3) 1723 (3)(3)(3)(3)(3)(3)(3)(3)(3)(3)(3)(3)(3)(84(8) 84] # 3 % 47710 %050-258 %010-258 % 010-258 % 00	2122 = = = = = = = = = = = = =	

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.

- 415	12								
148	21.76	R.8 8445225	-8-						
문서	표 사년 	도 비교법위 : (37)#4	·····	-					
		미달간 유사도비교		at the set of the	-				
		····································							
(0,\$		(日本2) 제작자(제품 (121)			+ 880	₩¥ 590건 (최대 유	1176-141-06-3		
96,74 19,42	均差 対策 月刊	년 4158 개봉 (262) 조합하지동 (1911) 봉 (168)			200 C	· 1468日 (山田 平小型			
96,74 19,4	対量 対量 月間	년 김대씨동 (262) 고입해제동 (191) 동 (163) 동 7(개봉 (141)	847	83.1	200 C			8224	
96,34 13,47 02,8	料整 料整 料理 料理	년 김대씨동 (262) 고입해제동 (191) 동 (163) 동 7(개봉 (141)		83 ÷		- 14662 (BLK RAG) + 2847 D+04	(40.35)		
96,34 19,47 02,30	料度 川間 月度	년205개동 (262) 고입하지동 (101) 동 (163) 동7(개동 (141) 구동동			4000 D+024	- 14662 (BLK RAG) + 2847 D+04	45.9%) 8419		
96,34 19,47 02,30	小型 小型 小型 二回	년 김대의 북 (262) 고입하지 등 (101) 등 (163) 등 7(제품 (141) 구동함			4565 Détta	- 14662 (BLK RAG) + 2847 D+04	84.9%)	(H112)	•
86,14 19,42 02,10	料理 川田 日 日 日	년 리15 세종 (262) 고립14 세종 (191) 동 (162) 동 71 세종 (141) 구동함 [* ************************************	80 Mb	4/200 D+014 2010/27000004	- 14662 (BLK RAG) + 2847 D+04	843 843 [201077700001.gs.100	N25	•
06,14 19,22 02,8 7		년 각대 제품 (262) 고립 체제 등 (101) 등 (163) 등 71제를 (141) 구동함 [고,방문 프로알에 제품 21_문 프 주글 알에 제품 21_문 프 주글 알에 주름	응한 후 전류강한성 감시 및 평가 사용표전로 유한감사	80.1% 57.3% 55.4% 53.0%	4/DHD Dema 4/DHD Dema 2010/27000004 2010/27000004 2010/27000004 2010/27000004	- 14662 (BLK RAG) + 2847 D+04	88.9%) E-419 2011077000001_2.8.100 2011077000004_2.8.100	(비원왕 (비원왕 (비원왕)	•
06,14 19,42 02,8 7 7		생각대체물 (262) 접실하지 등 (163) 률 (163) 률 (163) 률 (163) 로 (163) 지(21) 지(21) 지(21) 도 (163) 지(21) 도 (163) 지(21) 도 (163) 지(21) 도 (163) 지(21) 도 (163) (163	* 응한 후 전로간전성 감시 및 왕가 사용표전로 유한감사 전로입험제 계속권내 이용함 제거 고학전로 인용시 고치 사용표전로 인용 및 개장전	8 40 1% 57 3% 55 4% 51 0% 52 8%	41000 Devite 41000 Devite 2011073000004 20110737000004 20110737000004 20110737000004 20110737000004	- 14662 (BLK RAG) + 2847 D+04	84.5%) 84/5 0110727000001_08.6.100 00110727000001_08.4.100 00110727000001_08.100 0011073000004_08.100 001107300004_08.100	(비원왕 비원왕 비원왕 이 비원왕 이 비원왕	•
06,14 19,22 02,8 5 3 4 5 8		년210(비원 (US) 월25(14) 원 (US) 월 (143) 월 (145) 월 (145) 가방문(US) 가) 가방문(US) 가) 가방문(US) 가) 가방문(US) 가) 가) 가) 가) 가) 가) 가) 가) 가) 가) 가) 가) 가)	응한 중 전류전한성 감시 및 왕가 사용표전로 유한감사 전로 학왕 가 유권내 이 출동 차가 고학전로 인물시 고의 사용표전로 선물 및 가장전 인전로 인수 및 감사	80 1% 57 5% 53 4% 53 6% 52 5% 51 2%	4121212 D-6412	- 14662 (BLK RAG) + 2847 D+04	82.3%) 8413 2013/7200007_84 2013/7200007_84 2013/7200007_84 2013/7200007_84 2013/7200007_84 2013/7200007_84 2013/7200007_84 2013/7200007_84 2013/7200007_84 2013/7200007_84 2013/7200007_84 2013/7200007 2013/7200007 2013/7200007 2013/7200007 2013/7200007 2013/7200007 2013/72007 2013/72007 2013/72007 2013/720007 2013/720007 2013/720007 2013/720007 2013/720007 2013/720007 2013/720007 2013/720007 2013/720007 2013/720007 2013/720007 2013/720007 2013/7200007 2013/7200007 2013/72000000000000000000000000000000000000	(비파달 4 비파달 1 비파달 0 비파달 2 비파달 2 비파달	•
86,M 19,Q 19,Q 19,Q 12,B 1 2 3 4 5 9 7		실 2001 비용 (262) 전철해 위용 (103) 분 (163) 용 17지통 (141) 가지통 (141) 가지통 (141) 가지통 (143) 가지분 유민하여 가지는 문화 전망하여 가지는 문화 전망하여 가지는 문화 전망하여 가지는 문화 전망하여 가지는 문화 전망하여 가지는 문화 전망하여	응한 후 전류전전 5 고식 및 문가 사용 전문로 유진감사 전문 전문로 유진감사 전문 전문 사 유고 지수는 진을 시 규지 사용 지구는 전문 및 개장전 전문 진 단 ~ 문 관사 정전 전로 유금에 대한 우리 및 제항사항	40 1% 57 3% 55 4% 51 0% 52 5% 51 2% 51 2%	4421818 (2.4.8.8) 4421818 (2.4.8.8) 2410727000004 2410727000004 2410727000004 2410727000004 24110727000004 24110727000004 24110727000004	- 14662 (BLK RAG) + 2847 D+04	84.1%)	(비원왕 4 비원왕 1 비원왕 2 비원왕 2 비원왕 2 비원왕 2 비원왕	•
86,94 19,22 02,85 2 3 4 5 6 7 8		실 2001 비용 (202) 2014 14 % (101) 용 (104) 용 (104) 용 7/1 % (104) 7/8 % 8 % 1041) 7/8 % 8 % 1041) 7/8 % 8 % 1041) 7/8 % 8 % 1041 7/8 % 8 % 1041 7/8 % 8 % 1041 7/8 % 10	비원 비원 비원 비원 비원	40 V% 57 % 51 4% 51 4% 51 2% 51 2% 51 2% 51 2% 51 2% 51 2%	4101010 C+411 20110727000004 20110727000004 20110727000004 20110727000004 20110727000004 20110727000004 201107270000004 201107270000004 201107270000004 201107270000004	· 대표권 네네 주사간 · · · · · · · · · · · · · · · · · · ·	84.15) 8449 01137700001, () 8.00 01137700000, () 8.00 01137700000, () 8.00 01137700000, () 8.00 01137700000, () 8.00 011377000000, () 8.00 011377000000, () 8.00	(비원왕 4 비원왕 1 비원왕 2 비원왕 2 비원왕 2 비원왕 2 비원왕 1 비원왕	•
96, M 19, () 19, ()) 19, () 19, () 19, ()) 19, () 19, ()) 19,		성격대체용 (362) 전철체체용 (103) 분 (182) 분 (182) 분 (182) 분 (182) 전 (182) (182) 전 (182)	응한 중 전통한 등 감사 및 환자 승환 중 주요한 가 및 환자 사용호호류 수 있다 프로 등 환자 수 있다 프로 등 환	4 60% 57% 51% 51% 51% 51% 51% 51% 51% 51% 51%	4/10/20 0.0+07. 20110727000004 20110727000004 20110727000004 20110727000004 20110727000004 20110727000004 20110727000004 20110727000004 20110727000004 20110727000004	- 14662 (BLK RAG) + 2847 D+04	84.9%) 8449 24447700001, 28.40 24447700001, 28.40 24447700001, 28.40 24447700001, 28.40 24447700001, 28.40 24447200001, 28.40 2444720001, 28.40 24447200001, 28.40 244472000001, 28.40 244472000001, 28.40 244472000000000000000000000000000000000	비치왕 비치왕	•
56,94 19,22 02,35 7 2 3 4 5 8 7 8 9 7 8 9		실 2001 비용 (202) 2014 14 % (101) 용 (104) 용 (104) 용 7/1 % (104) 7/8 % 8 % 1041) 7/8 % 8 % 1041) 7/8 % 8 % 1041) 7/8 % 8 % 1041 7/8 % 8 % 1041 7/8 % 8 % 1041 7/8 % 10	비원 비원 비원 비원 비원	40 V% 57 % 51 4% 51 4% 51 2% 51 2% 51 2% 51 2% 51 2% 51 2%	4101010 C+411 20110727000004 20110727000004 20110727000004 20110727000004 20110727000004 20110727000004 201107270000004 201107270000004 201107270000004 201107270000004	· 대표권 네네 주사간 · · · · · · · · · · · · · · · · · · ·	84.15) 8449 01137700001, () 8.00 01137700000, () 8.00 01137700000, () 8.00 01137700000, () 8.00 01137700000, () 8.00 011377000000, () 8.00 011377000000, () 8.00	비치왕 비치왕	•

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.

312 4	2:4912/404932							
化效用 建制 化水合物								
2 15	492	843	8.45	사전관업 접수면 #	· · ::::::::::::::::::::::::::::::::::	2224	8934	8423
1	07,902#25121498	동안 등 만족강한성 값 시 및 분가	62.9%	201107210000091		201107270000091.년 로~1008 문전 중 연료건전성 검사 및 문가 (Trev. 00.7me)	NIME	
	67,424220438	전표합답에 계속한다 이 별질 제거	15,4%	261107270000081		201107270000097,전 로-(201 전로급함에 계측한 내 이물질 제가 (Rev.002.top)	1012	
2	07_박전로급입내처럼	PLUS7 Feel Ceston and Safety Eveluation for Korea Standard Nuclear Priver Plants	47,3%	201102210000014	201183280008228	281102210008014.jz.,359F- TR-DNR-0400149-A.adf	49	
1	07,9222224948	사용후면로 초등다갑사 (면로-1001)	43.3%	201204110000821		2012081100008021,면 로-1005 서문부언로 초음파 검사 (Rev. 003.5ve	100	
900 20	시 전작물자 412여 부 : 4	8 +						
12								

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 multilanguage 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