ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2014 September: 20(3): 109~131 http://dx.doi.org/10.13088/jiis.2014.20.3.109

사례 기반 지능형 수출통제 시스템 : 설계와 평가*

홍원의
한국과학기술원 지식서비스공학과
(laftworld@kaist.ac.kr)

김의현
한국과학기술원 지식서비스공학과
(uihyun@kaist.ac.kr)

조신희
한국과학기술원 지식서비스공학과
(chosinhee@kaist.ac.kr)

김산성
한국방송공사 기술연구소
(positivemd@gmail.com)

이문용
한국과학기술원 지식서비스공학과
(munyi@kaist.ac.kr)

신동훈
한국원자력통제기술원
(nucleo@kinac.re.kr)

최근 전 세계적인 원전 설비의 수요 증가로 원자력 전략물자 취급의 중요성이 높아지는 가운데, 국외 수출을 위한 원전 관련 물품 및 기술의 신청 또한 급증하는 추세이다. 전략물자 사전판정 업무는 통상 원자력 물자 관리에 해박한 전문가의 경험 및 지식에 근거하여 수행되어 왔지만, 급증하는 수요에 상응하는 전문 인력의 공급이 부족한 실정이다. 이러한 문제를 극복하기 위하여, 본 연구진은 전략물자 수출 통제를 위한 사례 기반 지능형 수출 통제 시스템을 설계 및 개발하였다. 이 시스템은 현장 전문가의 전담 업무이던 신규 사례에 대한 전략물자 사전판정 과정 업무의 주요 맥락을 자동화 하여 전문가 및 관계 기관이 감당해야 할 업무 부담을 줄이며, 빠르고 정확한 판정을 돕는 의사결정 지원 시스템의 역할을 맡는다. 개발된 시스템은 사례 기반 추론 (Case Based Reasoning) 방식에 기반을 두어 설계되었는데, 이는 과거 사례의 특성을 활용하여 신규 사례의 해법을 유추하는 추론 방법이다. 본 연구에서는 자연어로 작성된 전자문서 처리에 널리 사용되는 텍스트 마이닝 분석 기법을 원자력 분야에 특화된 형태로 응용하여 전략물자 수출통제 시스템을 설계하였다. 시스템 설계의 근거로 선행 연구에서 제안된 반자동식 핵심어 추출 방안의 성능을 보다 엄밀히 검증하였고, 추출된 핵심어로 신규 사례와 유사한 과거 사례를 추출하는 알고리즘을 제안하였다. 제안된 방안은 텍스트 마이닝 분야의 TF-IDF 방법 및 코사인 유사도 점수를 활용한 결과(α)와 원자력 분야에서 통용되는 개념적 지식을 계통으로 분류하여 도출한 결과(β)를 조합하여 최종 결과 (γ) 를 생성하게 된다. 세부 요소 기술의 성능 검증은 임상 데이터를 활용한 실험 및 실무 전문가의 의견수렴을 통해 이루어졌다. 개발된 시스템은 사전판정 전문 인력을 다수 양성하는 데 드는 비용을 절감하는 데 일조할 것이며, 지식서비스 산업의 의미 있는 응용 사례로서 관련 산업의 성장에 기여할 수 있을 것으로 보인다.

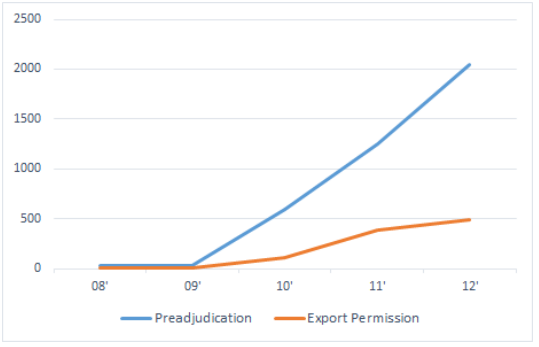
주제어 : 전문가 시스템, 수출 통제 시스템, 핵확산 방지, 사례 기반 추론

논문접수일 : 2014년 6월 29일 논문수정일 : 2014년 8월 3일 게재확정일 : 2014년 8월 24일
투고유형 : 국문급행 교신저자 : 이문용

1. 개요

전 세계적으로 원자력 시설 및 설비의 수요가 늘어남에 따라 전략 물자 수출 통제의 중요성이 대두되고 있다. 미국 9.11테러 사건을 계기로 유엔 (UN)은 대량 살상 무기 확산방지를 위하여 관련 원자력 연구, 개발, 생산, 사용, 수송 등의 용도로 이용될 수 있는 물자 및 기술과 같은 전략물자에 대한 통제체제를 구축할 것을 회원국에 요구하였으며, 북한과 이란의 핵 활동에 대해

* 본 논문은 한국원자력통제기술원(KINAC) 원자력안전위원회 재원(원자력안전연구사업)으로 지원된 연구임 (2013B3914004)



〈Figure 1〉 Change of nuclear preadjudication and export permission requests over time

서도 전략물자 수출 통제 규정을 강화하였다. 이와 같이 전략물자 수출 통제의 중요성이 대두되는 가운데 우리나라는 아랍에미리트 (UAE) 상용 원전 수출 및 요르단 연구로 수출을 성사시킴으로써, 국제사회는 우리나라의 전략물자 수출 통제 이행 현황을 예의 주시하고 있다.

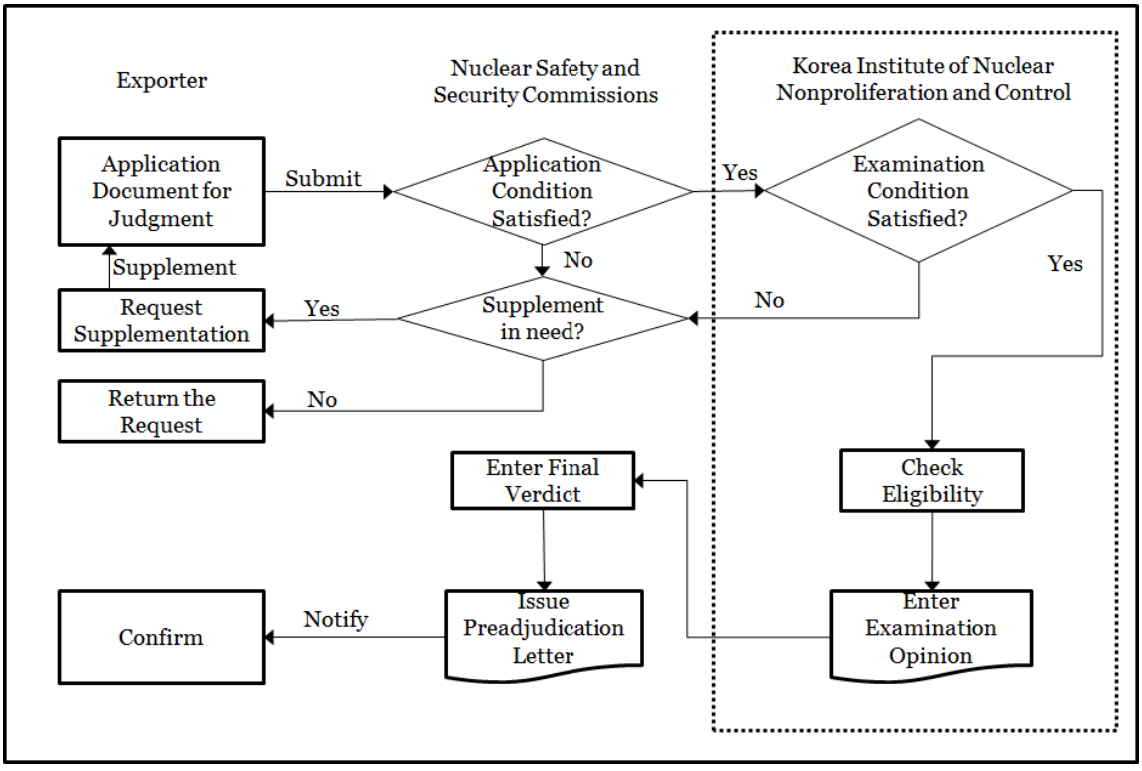
원자력 관련 수출 예정 물자는 원자력 통제 기술원의 사전 판정 및 수출 허가 과정을 철저히 준수하여 전략물자 비해당 여부가 확인된 사례일 경우에만 수출이 승인되며, 원자력 분야 전문가의 면밀한 분석 및 검토를 요구한다. 그러나 최근 들어 전략물자 해당여부를 검토 받아야 할 사례가 급격히 증가하고 있는 실정이다. UAE 원전 관련 수출물자의 경우 약 5,000여건 (보조기기 80,000여건)의 사전 판정 신청이 들어왔으며, <Figure 1>에 나타난 바와 같이 2010년 원전 수출 초기 대비 급격히 늘어난 사전 판정 신청건수로 정부, 관계기관 그리고 관련 기업의 부담이 커지고 있다. 우리나라는 현재 핀란드 등 약 10개국에 원전 건설 사업을 추진하고 있으며, 향후 우리나라가 4기의 원전 건설 사업을 수주한다면 약 20,000건 이상의 사전 판정 수행이 필요하게 될 것이다.[1] 따라서 대량의 사전 판정 및 수출 허가 신청 사례를 관리하고 나아가 심사 담당자의 의사결정을 지원할 수 있는 지능형 시스템 개발이 시급하다.

현 전략 물자 수출 통제 절차는 크게 사전 판정 절차와 수출 허가 절차로 나누어진다. 사전 판정은 <Figure 2>에서 볼 수 있듯 제조자, 수출업자가 취급하는 물품이 원자력 전용품목 및 기술에 해당하는지 여부를 확인하는 절차이며, 수출이 예상될 시 계약 상황 이전에 사전 판정을 신청하여야 한다. 수출 허가란 사전 판정 결과 전략 물자에 해당하는 원자력 전용 물품 또는 기술을 수출하는 경우에 원자력안전위원회 원자력통제팀에 신청하여 허가를 받는 과정을 말한다. 포괄적으로 수출 허가 절차는 사전 판정 절차를 포함하며, 사전 판정 절차는 전체 전략 물자 수출 통제 절차에서 가장 처리하기 어려운 부분이기 때문에, 본 연구진은 주로 사전 판정 절차를 자동화 하는 데 중점을 두고 연구를 진행하였다.

사전 판정 절차에서 취급 품목 및 기술이 전략물자에 해당되는지 여부는 한국원자력통제기술원(KINAC)의 전문가들이 검토하게 된다. 이때 물품 매뉴얼, 상품안내서 또는 사양서 등 수출품목의 성능과 용도 및 수출대상 기술의 내용이 표기된 서류 등이 제출되어야 하는데, 이러한 절차는 사전 판정, 수출 허가, 핵 물질 수출입 승인 및 보고 등 관련 업무를 소개하고 실제 업무 수행을 도와주는 NEPS 웹사이트 원자력 수출입 종합지원 시스템[2] 을 통해 이루어진다.

제출된 서류들은 심사 담당자의 사전 판정 절차를 거치게 되며, 이때 담당자는 관계법령, NSG (Nuclear Suppliers Group) 핸드북, 심사지

1) 한국원자력통제기술원: www.kinac.re.kr
2) NEPS: http://www.neps.go.kr/



〈Figure 2〉 Preadjudication Process

침서, 과거판정사례 데이터베이스 등의 자료를 근거로 원자력 분야 전문가로서의 종합적인 견해를 제시하고 이를 통해 최종 판정을 내리게 된다. 하지만 현재까지 이러한 심사 절차는 주로 전문가의 암묵적 지식 (implicit knowledge)과 개인적 경험에 의존하여 수행되어 왔기 때문에 업무 처리 속도가 제한적이라는 한계가 있었다. 따라서 급증하는 전략물자 수출입 통제 수요에 보다 효과적으로 대처하기 위해서는 전문 심사관들의 축적된 지식과 노하우 및 컴퓨터의 정보처리 능력과 추론, 그리고 기계학습 능력을 결합한 지능형 수출입 통제 기술이 요구된다.

이와 같이 지능형 수출 통제 시스템 개발 연구의 최종 목적은 기하급수적으로 증가하는 사전 판정 신청 건에 대해 보다 신속한 판정을 내릴 수 있도록 심사 담당자의 의사 결정을 지원할 수 있는 시스템을 개발하는 것이다. 이를 위하여 현장 전문가인 심사 담당자의 의사 결정 과정과 현재까지 누적되어 온 과거 판정 사례의 활용, 그리고 각종 법령 및 절차적 기준을 유기적으로 연결하는 통합 시스템을 구상할 필요가 있다.

본 연구진은 수많은 사전 판정을 효율적으로 처리하기 위한 지능형 수출 통제 시스템의 설계 의뢰를 한국원자력통제기술원 수출입 통제 팀으로부터 요청 받아 본 프로젝트를 수행하였다. 총 연구기간 23개월간 과거의 사례 데이터베이스로부터 신청 사례와 유사한 사례를 발견해 내고, 그로부터 새로운 사례의 판정 실마리를 찾아내

는 사례 기반 추론 시스템을 개발하였다.

신규 사전 판정 건으로 처리되는 기술 혹은 물품은 그를 상세히 설명하는 문서(설계도, 사용설명서 등)들로 표현된다. 사례 기반 추론 시스템은 이전에 심사 담당 전문가가 직접 전략물자 해당 혹은 비해당 판정을 내렸던 사례를 활용하여 신규문서의 사전 판정 결과를 추론하는 시스템이며, 본 연구에서는 일반적으로 널리 활용되는 텍스트 마이닝 기법에 기반을 두어 이를 원자력 분야에 특화시키기 위한 방안을 연구하였다.

사례 기반 추론 시스템의 유효성 평가를 위해서 본 연구진은 임상 데이터를 활용한 실험 및 성능검증을 시행하였고 개선사항을 반영한 추가 실험을 완료하였으며, 과제 수행기간 중 위탁업체와의 정기적인 의견교류를 바탕으로 사전 판정 심사관과의 피드백을 종합적으로 수집해 활용하였다.

본 연구를 통해 개발한 추론 기술들은 지능형 전략 물자 수출 통제 시스템의 핵심에 해당하는 부분이며, 이를 활용하여 구현된 시스템은 전략물자 판정 및 심사에 소요되는 기간을 단축시켜 이에 필요한 경제적, 행정적 부담을 최소화하고 보다 체계적이며 객관적인 심사결과를 도출해 낼 수 있을 것으로 기대한다.

2. 관련 연구

2.1 전문가 시스템

전문가 시스템 (Expert System)이란 특정 분야의 현장 전문가로부터 얻어진 전문 지식을 구체적 논리로 표현하여 만들어진 지식 기반의 시스템을 말하며, 전문가와 비슷한 수준의 추론을 통해 주어진 문제에 대해 조언하고 결정을 돕는 역할을 하는 지능적 의사결정 지원 시스템이라 할 수 있다 (Kendal and Creen, 2007). 전문가 시스템은 AI의 여러 분야에서 전문가의 업무 수행 지식을 컴퓨터로 옮기는 작업을 기본으로 발전되어 왔다 (Liao, 2005).

본 연구진에서 설계한 지능형 전략물자 수출통제 시스템은 이러한 전문가의 문제해결에 관한 작업 능률과 신뢰성 향상을 목적으로 하고 있다. 이는 수많은 사전 판정 과정에 수반되는 업무들을 보다 신속하고 정확하게 처리하기 위한 시스템으로 전문가의 최종결정에 도움을 준다. 본 연구를 통해 개발한 시스템은 사례 기반 추론 방식을 기반으로 설계되었으며, 이 방식을 통해 얻을 수 있는 장점을 최대한 활용할 수 있도록 시스템을 설계하였다.

2.2 사례 기반 추론

사례 기반 추론이란, 인간이 경험을 통해 얻어진 지식을 가지고 문제를 해결해 나가는 것과 같은 방식의 추론 방법이다 (Kendal and Creen, 2007). 즉, 인간의 지적 활동을 모델화한 것으로 과거 문제로부터 얻은 상황 경험이나 지식을 사례 데이터베이스로 구축하여 어떠한 상황이나 문제가 발생하면 기존의 사례 데이터베이스에서 가장 똑같거나 또는 가장 유사한 사례를 선택하여 그 사례가 제시하는 해결책으로 현 문제에 대한 답을 제시한다 (Lee, 1996).

사례 기반 추론은 과거의 사례를 바탕으로 문제를 해결하기 때문에 비록 문제가 복잡하더라도 이미 해결된 사례를 통해 신속히 해를 도출할 수 있다. 그러므로 지식이 잘 파악되지 않는 대상영역에 있어서도 사례로서 추론을 가능하게

한다. 그리고 정확히 일치되는 사례를 발견할 수 없다면 가장 유사한 사례를 변형하여 새로운 문제를 해결하도록 할 수 있으며 이렇게 해결된 사례는 다시 새로운 사례로서 저장되게 된다. 그러나 사례 기반 추론은 일반적인 상식에 대해서 쉽게 표현이 되지 않으며, 도출된 결론에 대해 설명이 어렵다. 또한 충분히 많은 사례가 있는 경우에는 지식 습득에 문제가 되지 않지만, 그렇지 못할 경우 문제가 된다 (Prentzas and Hatzilygeroudis, 2007).

사례 기반 추론 시스템은 전문가들로부터 경험적 사례를 추출하여 축적하는데, 새로운 판단이 필요할 경우 검색이 용이하도록 사례 베이스를 체계적으로 구축해야 한다. 사례 기반 추론을 실행하는 데 있어 쟁점이 되는 것은 사례 간 유사도의 정의인데, 여기에는 범용적인 정의가 없기 때문에 적용 분야 또는 개발하는 시스템의 특성을 반영하여 결정해야 한다.

2.3 텍스트 마이닝

텍스트 마이닝 (Text Mining) 기술은 지식 집약형 프로세스이며, 비정형 텍스트 데이터 (Unstructured Text Data)로부터 가치 또는 의미 있는 정보를 추출해내는 기술이다 (Feldman, R. and Sanger, 2007). 즉, 텍스트 마이닝은 데이터 마이닝의 한 분야로 방대한 텍스트 데이터를 기반으로 의미 있는 패턴을 찾아내는 기술이다 (Navathe and Elmasri, 2000). 이것은 텍스트로부터 지식을 추출하는 알고리즘 및 자연 언어 처리 (Natural Language Processing) 기술에 기반하고 있다. 인간의 말은 각 언어별로 어휘적, 문법적 독특성이 있을 뿐 아니라, 그 표현의 형태가 매우 다양하고 복잡하여 일괄적인 규칙으로 규정하기 힘든 경우가 많으며, 언어가 사용되는 환경에 따라 끊임없이 변화하는 특성을 지니고 있다. 이러한 인간 언어 중 문자로 표현된 언어를 컴퓨터로 분석 처리하고 그 구조와 의미를 이해하고자 하는 기술이 자연언어처리 기술이다 (Kodratoff, 1999). 컴퓨터가 개발되면서부터 끊임없이 연구되어 온 분야이지만, 언어가 가진 복잡성 때문에 아직도 도전적 목표가 많이 남아 있는 기술 분야로 손꼽힌다.

사용자는 이러한 텍스트 마이닝 기술을 통해 수많은 정보 사이에 의미 있는 정보를 추출해 내고, 다른 정보와의 연계성을 파악하며, 텍스트가 가진 카테고리를 찾아내는 등, 단순한 정보 검색 이상의 결과를 얻어낼 수 있다. 컴퓨터가 인간이 사용하는 언어로 기술된 정보를 깊이 분석하고 그 안에 숨겨진 정보를 알아내기 위해서는 대용량 언어자원과 복잡한 통계적, 규칙적 알고리즘이 적용되어야만 한다. 컴퓨터와 인간의 언어 사이에 가지는 차이가 매우 크지만, 많은 기술 분야에서 이를 좁히려는 발전이 이루어져 왔다 (Gupta and Lehal, 2009).

본 시스템에서는 수많은 문서들로부터 심사에 관련된 의미 있는 정보를 추출해 심사관에게 제공하기 위해 텍스트 마이닝 기법을 활용한다. 따라서 이와 관련된 텍스트 마이닝 기술을 문헌 연구하였으며, 주요 기술 분야는 다음과 같다.

1) 정보추출 (Information Extraction)
2) 문서분류 (Document Categorization)
3) 문서군집 (Document Clustering)

첫째, 정보추출 (Information Extraction)은 텍스트 문서 내에서 중요한 의미를 가지는 정보들을 자동으로 추출하는 기술이다 (Gupta and Lehal,

2009). 사용자는 정보추출 기술을 통해, 비정형 문서에서 중요 키워드, 핵심 개념, 특정 사건, 인명, 지명, 날짜, 상황 및 조건, 결론 등의 다양한 정형 정보를 추출하여 활용할 수 있도록 돕는다. 키워드와 같은 기본적인 정보는 자동 분류, 군집 등에 직접적으로 활용되는 중요 요소가 되고, 그 외의 다양한 상세 정보들은 자동 요약에 있어서 매우 중요한 문장 구성 요소가 된다. 최근 들어서, 정보추출 기술은 경쟁자 정보 분석, 조직 내의 위험 관리 시스템 개발, 온톨로지 기반의 시맨틱 웹 기술을 구현할 때, 비정형 텍스트 문서에 의미 정보를 부착하는 기술의 개발, 그리고 기존 정보 시스템의 성능을 개선하여 효과적인 정보 접근 및 관리를 가능하게 하는 핵심 기술로 각광을 받고 있다.

둘째, 문서분류 (Document Categorization)는 정의되어 있는 주제 또는 분류체계에 따라 문서를 분류하는 기술이다 (Gupta and Lehal, 2009). 기술의 발전과 인터넷의 활성화는 엄청난 정보의 생산과 유통을 가능케 했으며, 이로 인해 분산되어 상호 복잡하게 연계되어있는 방대한 정보를 분류하기란 매우 어려운 과제가 되었다. 문서분류 시스템에서 컴퓨터를 이용 문서에 대해 분류를 할 때는 단순히 유사한 의미의 단어 수만 세는 것이 아니라, 전체적인 문서의 내용 및 주제에 맞게 분류한다. 이를 위해 문서분류 시스템은 주제, 관련성, 또는 단어들의 동의어, 관련어, 그리고 광범위한 또는 좁은 의미 등에 대해 정의해 놓은 유의어 사전을 미리 만들어 이용하기도 한다.

셋째, 문서군집 (Document Clustering)은 비슷한 문서들끼리 묶어주는 기술이지만, 문서분류는 미리 정의된 주제 또는 분류체계 안에서 분류한다는 점에서 다르다 (Liritano and Ruffolo, 2001). 문서군집은 각 지식 콘텐트의 특성을 파악해 그 내용 혹은 형태가 유사하거나 상호 관련성이 높은 콘텐트들을 군집시켜 주는 기술이다. 사용자는 문서군집 기술을 통해, 관심 있는 문서들을 그 관련도 순으로 한꺼번에 묶어서 효과적으로 검토해 볼 수 있을 뿐만 아니라, 통상의 문서군집 기술은 대상 문서의 언어학적 분석을 통해 차별화된 중요 특성들을 추출해 내고, 이를 다른 문서의 특성들과의 비교 (유사도 계산)를 통해 그 유사도가 높은 문서들을 상호 묶어주는 방식으로 구현된다. 정확한 유사도의 계산과 효과적인 군집을 위해 다양한 통계 기반, 규칙 기반 알고리즘들이 연구되어 왔다.

2.3.1 텍스트 마이닝 연구현황

디지털화된 방대한 양의 문서를 처리하는 텍스트 마이닝의 연구는 오래 전부터 이루어져 왔다. <Table 1>은 최근 5년간의 주요 연구 및 응용분야를 보여준다. 텍스트 마이닝 기술은 의학, 법학, 과학 등 사회 전반에 걸쳐 다양한 분야에 적용되어 왔음을 알 수 있다. 본 연구에서는 원자력이라는 특정 분야에 대한 전문지식을 이용하고 있다. 이는 매우 한정적이며, 이러한 전문지식에 대하여 텍스트 마이닝을 활용한 시스템은 현재 개발되어 있지 않다. 본 연구원들은 원자력이라는 특정 전문지식을 타겟으로 한 텍스트 마이닝 기술을 이용하여 지능형 전략물자 수출통제 시스템에 필요한 알고리즘을 설계하는 작업을 진행해 왔다.

2.3.2 텍스트 마이닝 알고리즘

다량의 전자문서로부터 의미를 추출하는 텍스트 마이닝 알고리즘에는 여러 가지 방식이 알려

〈Table 2〉 Recent 5-year Summary of Research in Text Mining (2009~2013)

Application Area	Study
Bibliography	Vellay et al., 2009; Liu et al., 2010
Biology	Krallinger et al., 2009; Krallinger et al., 2010
Chemistry	Jessop et al., 2011
Decision Support	Rajpathak et al., 2012
Education	Lin et al., 2009; Hung, 2012
Information Retrieval	Li and Wu, 2010
Law	Wyner et al., 2010; Firdhous, 2012; Chen et al., 2013
Management	Netzer et al., 2012; Yoon, 2012
Material Engineering	Lee et al., 2013
Medicine	Hur et al., 2009; Yang et al., 2009; Al-Mubaid and Singh, 2010; Kozomara and Griffiths-Jones, 2011; Landeghem, 2011; Ananiadou et al., 2013; Rak et al., 2012; Xie et al., 2013
Music	Hu et al., 2009
Social Media	Corley et al., 2010
Social Network	Macskassy, 2011
Social Review	Ananiadou et al., 2009; Cao et al., 2011; Ghose, 2011

져 있는데, 자동 키워드 추출은 전자문서의 특징을 잘 기술하는 핵심어들의 집합을 얻는 과정이다 (Hulth, 2003). 문서의 특징을 잘 기술하는 핵심어를 일일이 문서를 읽으면서 추출하는 것은 시간과 노력이 많이 소요되는 과정이다. 따라서 자동 핵심어 추출을 위한 여러 가지 방안이 제시되어 왔다 (Aizawa, 2003; Hulth, 2003; Yan et al., 2013).

보편적으로 널리 사용되는 방안 중 하나로 TF-IDF 방식이 있다. TF-IDF (Term Frequency - Inverse Document Frequency)는 정보 검색과 텍스트 마이닝에서 이용하는 측정지표로, 여러 문서로 이루어진 문서군이 있을 때 어떤 단어가 특정 문서 내에서 얼마나 중요한 것인지를 나타내는 통계적 수치이다. 문서의 핵심어를 추출하거나, 검색 엔진에서 검색 결과의 순위를 결정하거나, 문서들 사이의 비슷한 정도를 구하는 등의 용도로 사용할 수 있다. TF-IDF는 기본적으로 어떤 단어가 문서에서 얼마나 자주 발견되는지를 기반으로 계산된다. TF (Term Frequency)는 특정한 단어가 문서 내에 얼마나 자주 등장하는지를 나타내는 값으로, 이 값이 높을수록 해당 단어는 문서에서 중요하다고 생각할 수 있다. 하지만 그 해당 단어가 문서군 내에서 자주 사용되는 경우, 이것은 그 단어가 그 문서군 내에서는 흔하게 등장한다는 것을, 다시 말하면 문서군 내에서 각 문서의 고유한 특징을 나타내지 못한다는 것을 의미한다. 이것을 DF (Document Frequency)라고 하며, 이 값의 역수를 IDF (Inverse Document Frequency)라고 한다. TF-IDF는 TF와 IDF를 곱

한 값이다.

<Table 2>의 내용은 TF 값은 특정 문서에서 자주 발견되는 단어는 적게 발견되는 단어보다 더 중요하다는 의미를 가진다 (Aizawa, 2003). 반면 IDF 값은 전체 문서 집합에서 적게 발견되는 정도를 의미한다 (Aizawa, 2003). IDF 값은 문서 집합 내의 총 문서 수를 그 단어를 포함하는 문서 개수로 나눠서 얻어진다. 예를 들어, 'nuclear'라는 단어가 어떤 문서에서 자주 발견되었다 하더라도, 해당 문서 집단 내의 대부분 문서에서 발견되는 흔한 단어인 경우, 이 단어는 특정 문서를 대표하는 중요한 단어라고 볼 수 없다. 따라서 IDF 값은 문서군의 성격에 따라 결정되게 된다.

〈Table 2〉 TF-IDF Definition

TF	$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$ $n_{i,j}$: *the number of occurrence of i in j* $\sum_k n_{k,j}$: *the number of occurrences of all term in document* d_j
IDF	$IDF_i = \log \frac{\lvert D \rvert}{\lvert \{d_j : t_i \in d_j\} \rvert}$ $\lvert D \rvert$: *total number of documents* $\lvert \{d_j : t_j \in d_j\} \rvert$: *number of documents where the term appears*
$TF-IDF$	$TF-IDF_{i,j} = TF_{i,j} \times IDF_j$

2.3.3 알고리즘 평가기준

정보검색 (Information Retrieval) 분야에서 Precision과 Recall은 검색 결과로 나온 문서 집합 (Set of Retrieved Documents)과 실제로 관련이 있는 문서 (Set of Relevant Documents)의 개수를 통해 정의된다. 따라서 일반적으로 Precision은 검색된 문서들 중 실제로 관련이 있는 문서들의 비율로 정의되며, Recall은 실제로 관련이 있는 문서들 중 검색 결과로 제시된 문서들의 비율로 정의된다.

텍스트 마이닝 분야의 관점에서도 문서 분류 결과의 성능평가는 Precision, Recall, F-measure를 통해 확인한다 (Yang, 1999). Table 3에 정리된 것과 같이, 이 문맥에서의 Precision은 정답으로 분류한 문서들 (True Positive + False Positive) 중 실제 정답인 문서들 (True Positive)의 개수를 의미하며, Recall은 실제 정답인 문서들 (True Positive + False Negative) 중 정답으로 분류한 문서들 (True Positive)의 비율을 의미하며, F-measure는 Precision과 Recall을 동등한 가중치로 계산한 조화평균을 의미한다.

〈Table 3〉 Precision, Recall, and F-measure

Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
$F-measure$	$\frac{2 \times Precision \times Recall}{Preciaion + Recall}$

* TP = True Positive (classified to be relevant (positive) when actually relevant), FP = False Positive (classified to be relevant when actually irrelevant), FN = False Negative (classified to be irrelevant when actually relevant)

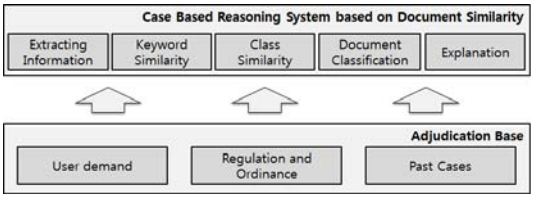
본 연구에서는, 핵심어로 대표되는 각 계통에 가장 잘 속하는 문서를 찾는 것을 목적으로 한다. 따라서 지금까지 설명한 Precision 대신

Precision at n이라는 측정지표를 사용하게 되는데, Precision이 분류기가 내놓은 모든 답을 정답으로 취하는 데 비하여, Precision at n은 정답이라고 시스템으로부터 주어진 결과 중 상위 점수 n개의 결과만을 끊어서 고려하게 된다 (Powers, 2011). 본 실험에서는 Precision at 1을 사용하였는데, 이는 한 문서를 가장 적합한 하나의 계통에 할당하겠다는 의미이다.

3. 사례 기반 추론 방법론

본 장에서는 심사관의 전략물자 사전 판정을 신속하고 정확하게하기 위한 시스템 개발 기술의 개념 및 연구현황을 기술하였으며, 시스템 개발에 주로 사용된 텍스트 마이닝 알고리즘 및 성능평가 기준에 대하여 설명한다. 본 시스템에 적용되는 기술들은 현재 다양한 분야에서 적용되고 있으며, 꾸준히 연구가 진행 중에 있다. 앞서 기술한 바와 같이, 원자력이라는 특정 전문지식에 맞는 시스템 개발연구는 매우 미흡한 상황이며, 본 연구에서는 사례 기반 추론을 활용한 전문가 시스템 (Expert System)을 텍스트 마이닝 (Text Mining) 알고리즘에 기반을 두어 설계하였다.

<Figure 3>와 같이 사례 기반 추론 시스템은 신규문서의 내용과 유사한 과거 문서 자료를 수집하고 문서의 특징을 잘 반영하는 속성을 선택한 후, 과거 문서들의 사전판정 정보를 바탕으로 신규 문서의 사전판정 결과를 도출한다. 이러한 시스템의 설계를 위해서는 양질의 정보를 담고 있는 과거 사례 및 사례의 특징을 명시적으로 표현하는 단계가 요구되며 이를 위해서는 문서대문서 핵심어 기반 유사도 분석 기술이 필요하다.



〈Figure 3〉 Case Based Reasoning System Framework

텍스트 마이닝 분야에서 기초적으로 활용되는 문서 유사도 분석 알고리즘으로는 코사인 유사도 (Cosine Similarity) 방식이 있다. 코사인 유사도 방식에 따른 전략물자 해당/비해당 판정 과정은 다음과 같다.

1) 사전 판정을 내려야 할 신규문서가 입력되면 신규문서와 과거문서의 코사인 유사도를 비교한다.
2) 유사도가 높은 순서대로 과거문서를 정렬하고 상위 3건의 문서를 선정한다.
3) 선정된 문서의 유사도를 가중치로 활용하여 해당/비해당에 따른 가중평균 (weighted sum)을 계산함으로써 최종 점수를 산출하게 된다.

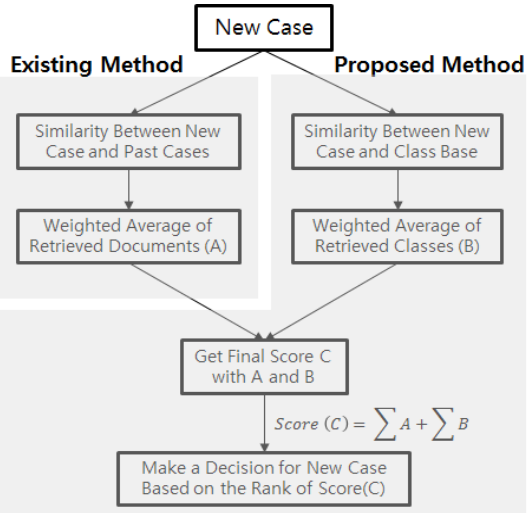
그러나 이와 같은 단순 핵심어 기반 비교 추론 과정은 원자력 분야처럼 그 범위가 특수하게 제한되어 있는 상황을 모사하는데 부족하다. 이처럼 문서의 문맥을 고려하지 않은 핵심어 기반 추출방식은 특정 분야의 개념적 특성을 반영한 문서분류가 어렵다는 단점이 있으며, 특히 원자력 분야에서는 핵심어 기반 문서 유사도보다 문서를 포괄하는 계통의 개념이 오히려 결과에 지배적인 영향을 미칠 수도 있다. 따라서 본 연구는 기존 문서가 가진 정보를 분석하여 정보를 추출한 후, 문서 비교 유사도 알고리즘과 계통 비교

유사도 알고리즘을 통합하여 결과를 도출하는 전략물자 수출 사전 판정 추론 시스템의 프레임워크를 아래 <Figure 4>와 같이 수립하고자 한다.

제안한 사례 기반 추론 시스템의 프레임워크에서는 기존의 문서 대 문서의 유사도를 비교하던 통념을 포함하는데, 추가적으로 신규 문서와 계통의 유사도를 고려하여 문서의 의미적인 측면을 반영한 결과를 내는 것을 목표로 한다. 이를 위해서 원자력 분야의 계통을 명시적으로 정의해야 할 필요성이 대두되는데, 연구진은 각 계통을 대표하는 핵심어를 선정하여 이 문제를 해결하고자 하였다. <Figure 4>의 각 단계를 설명하자면 다음과 같다.

1) 각 계통의 대표 핵심어를 정리해 계통 기반(class base)을 준비한다.
2) 전략물자 해당 또는 비해당 판정을 내려야 할 신규문서가 입력되면 계통을 대표하는 핵심어 집합내 단어들을 신규 문서와 비교하고, 핵심어가 신규문서에서 많이 검색되는 순서대로 상위 3개의 계통(retrieved class)을 선정한다.
3) 선정된 계통과 신규문서의 유사도를 가중치로 활용하여 해당/비해당에 따른 가중평균을 계산한다.
4) <Figure 4>에서 보듯이, 최종 점수 (C)는 핵심어 기반 계산 결과 (A)와 계통 기반 계산 결과(B)를 합하여 산출하게 된다.

본 장에서는 새롭게 제안한 프레임워크를 설계하는 데 기반이 되는 (1) 반자동 방식의 원자력 계통 핵심어 추출 방법과 (2) 원자력 계통 정보를 활용한 사례 기반 전자문서 분류 방법을 설명한다. 이어서 이를 기반으로 개발된 프로그램의 구동 결과를 예시로 제시하였다.



〈Figure 4〉 Existing vs. Proposed Document Analysis Method

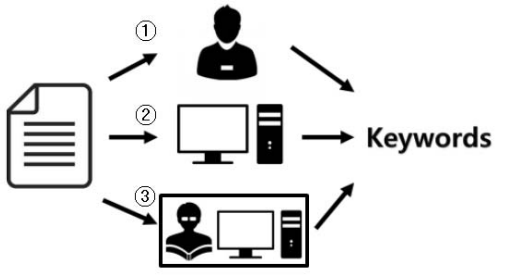
3.1 원자력 계통 정보를 활용한 사례 기반 전자문서 분류 방법

3.1.1 원자력 계통 핵심어 추출 방법

문서 분류를 수행하기 전, 문서의 특징을 대표하는 핵심어 추출이 선행되어야 한다. 문서로부터 핵심어를 추출하는 방안으로는 <Figure 5>에서 볼 수 있는 바와 같이 ① 완전 수동식 (Full-manual) 혹은 ② 완전 자동식 (Full-automatic) 방식이 사용되어 왔다. 완전 수동식 핵심어 방식은 추출 시간이 지나치게 오래 소요된다는 단점이 있고, 결과에 사람의 주관적 편향성이 반영될 수 있으며, 정확한 판단을 내릴 수 있는 전문가가 완성되기까지 오랜 세월을 거친 경험을 요구한다는 한계가 있다. 반면에 완전 자동식 핵심어 추출 결과는 분석 시간이 적게 걸리나 문서의 의미적인 특징을 제대로 반영하지 못한다는 단점이 있다. 따라서 <Figure 5>에서와 같이 두 방식을 절충하여, 한번 TF-IDF방식으로 핵심어를 추

출한 후 추출된 결과를 전문가가 다시 읽고 정제하는 ③ 반자동식 (Semi-automatic) 핵심어 추출 방식을 본 연구에 도입하였다 (Kim et al., 2014). 반자동식 방식은 두 번에 걸쳐 실험이 진행되었으며, 실험 1에서는 원자력 공학을 전공하는 학생들이 전문가의 역할을, 실험 2에서는 다년간 현장에서 심사업무를 수행한 현장 심사관이 전문가의 역할을 수행하였다.

이러한 세 가지 상이한 방법은 총 46개의 사전판정 신청 문서에 각각 적용되었으며, 그 결과 각 방식마다 각 문서에서 5개의 핵심어들이 추출되었다. 이어서 문서에서 추출된 핵심어들과 원자력 134개의 계통을 설명하는 문서들 간의 핵심어가 비교되었으며, 핵심어들 간의 매칭이 가장 많은 1개의 계통(Precision at 1)이 해당 문서의 정답으로 선택되었다. 최종적으로 이러한 기계적 분류는 현장에서 사전판정 업무를 담당해온 실무 전문가가 46개의 문서들 각각에 대해 미리 정해 놓은 정답과 비교하여 분류의 정확성이 평가되었다.



〈Figure 5〉 Three Keyword Extraction Methods

서로 다른 핵심어 추출 방법에 대한 실험결과는 다음 <Table 4>에 나타난 바와 같이, 제안하는 반자동식 핵심어 추출 방법이 보다 나은 성능을 보임을 알 수 있다. 예를 들어 Precision at 1 결과를 보면, 두 가지 반자동식 핵심어 추출방식 실험 결과의 평균값은 수동식보다 17.8%, 자동식보다 38.1% 향상되었다.

〈Table 5〉 Results of Keyword Extraction Experiment

	Extant Method		*Proposed Method*	
	Full Manual	*Full Automatic*	*Nuclear-major Students (1st experiment)*	*Field Expert (2nd experiment)*
Precision at 1	*0.434*	*0.370*	*0.500*	*0522*
Recall	*0.426*	*0.362*	*0.489*	*0511*
F-measure	*0.430*	*0.366*	*0.495*	*0.516*

Kim et al. (2014)의 연구에서는 원자력 분야 현장 전문가의 수동식 추출 방법과 TF-IDF 기반의 자동식 추출 방법에 비해 원자력공학 전공 학부 학생들의 반자동식 추출 방법이 핵심어 추출에 효과적임을 검증하였는데, 보다 엄밀한 비교를 위해서 반자동식 추출 방법을 수행하는 주체 또한 원자력 분야의 전문가가 되어야 논거가 보다 타당할 것으로 판단하여 추가 실험을 실시하였고 이 결과는 <Table 5>에 나타나 있다. 결과는 근소한 차이이기는 하나 현장 전문가의 반자동식 추출 방법이 전공 학생의 그것보다 효과적인 것으로 나타났으며, 두 집단의 차이가 근소하다는 점에서 반자동식 추출 방법이 수행 주체에 크게 의존하지 않는 유연한 방식이라는 것 또한 이해할 수 있다.

실험 결과를 요약하자면, Kim et al. (2014)에서 제안한 반자동식 핵심어 추출 방법은 사용자의 전문성에 크게 의존하지 않는 결과로, 단지

그 정확도에서 뿐 아니라 핵심어 추출에 소요되는 시간을 수동식 추출 방법에 비해 상당히 줄이는 것이며 이는 장시간의 수동 추출 과정에서 발생할 수 있는 편향성 및 오류의 가능성을 줄일 수 있을 것으로 보여진다.

<Table 4>에 기술된 결과는 사람의 개입이 전혀 없이 문서를 분류한다는 가정 하에서 Precision at 1을 사용하여 얻은 것이다. 이러한 조건에서 Precision은 최대 0.522의 성능을 보이고 있는데 실무적 환경에서는 여러 개의 후보들 중에 사람이 개입하여 1개의 계통을 최종적으로 고르게 되므로 n의 값은 1보다 높을 수 있으며, 이러한 실무적 환경에서의 성능은 <Table 4>에 나타난 값 보다 훨씬 더 향상될 것이다.

3.1.2 문서 유사도 계산 방법

문서-문서 유사도를 계산하기 위한 TF-IDF 와 코사인 유사도 방식으로는 여러 문서가 공통으로 가지는 특징을 반영할 수 없다. 특히, 원자력 분야의 개념적 특성을 반영한 문서분류가 어렵다는 단점이 있다. 따라서 본 연구는 문서-문서 유사도를 이용한 분류 기법에 더하여, 여러 문서를 계통으로 묶어 원자력 분야 문서의 계통적 특성을 고려한 문서 분류 기법을 제안한다. 본 과정은 크게 세 과정으로 나누어 볼 수 있다. 문서대 문서의 유사도 점수 (α)를 계산하는 과정과 문서대 계통의 유사도 (β)를 계산하는 과정, 그리고 최종 점수 (γ) 및 판정 결과를 제시하는 과정이다. 각 과정에 대한 상세 내용은 다음과 같다.

문서대 문서의 유사도 점수 (α) 계산

1) 하나의 신규문서를 입력받는다.
2) 기존 문서 각각에 대하여 TF-IDF방식으로 유사도를 계산한다.
3) 기존 문서들에 대하여 해당이면 1, 비해당이면 -1의 가중치를 부여한다.
4) TF-IDF점수가 0.5 이상인 문서들 각각이 가진 TF-IDF값들의 가중평균을 계산한다.

문서대 계통의 유사도 (β) 계산

1) 총 24개 계통 각각에 할당된 핵심 키워드 중 몇 개가 신규문서 내에서 발견되는지 검색한다.
2) (1)의 검색 건수에 대한 내림차순으로 24개 계통의 순위를 매긴다.
3) (2)에서 상위 3개 계통을 선정한다.
4) 계통의 해당 확률 각각에 상대적 키워드 개수를 곱하여 가중평균을 계산한다.

최종 점수 (γ) 및 해당/비해당 판정식

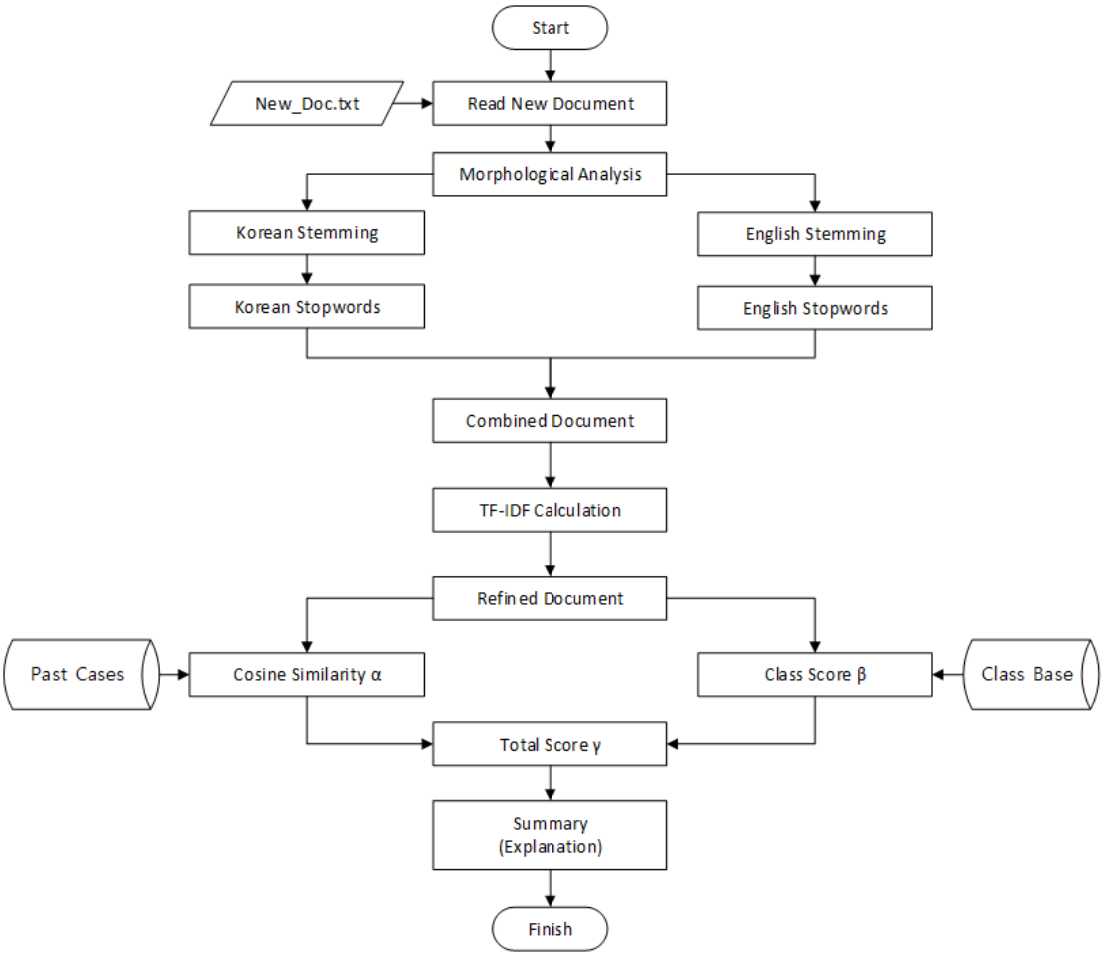
1과 2에서 구한 α 및 β에 대하여 최종 점수 γ를 다음과 같이 계산한다.

1) $If\ \alpha\beta > 0\ then\ \gamma = \frac{(\alpha+\beta)}{2} * 100$
2) $If\ \alpha\beta < 0\ then\ \gamma = (\alpha+\beta)*100$
3) $If\ \alpha\beta = 0\ then\ \gamma = 0$ [3]
4) $If\ \gamma \geq 0\ then\ [\text{해당}]\ \text{for}\ \gamma\%$
5) $If\ \gamma < 0\ then\ [\text{비해당}]\ \text{for}\ |\gamma|\%$

4. 사례 기반 추론 시스템 개발

사례 기반 추론 시스템을 개발하기 위해서는

3) γ=0 인 경우는 해당/비해당 판정을 내리기 어려운 신뢰도이나, 본 시스템의 문맥을 고려했을 때, 기본값 (Default)으로 '전략물자 해당' 판정을 내려야 기본적으로 수출을 제한할 수 있다는 데서 타당성을 가짐



〈Figure 6〉 Flowchart of Case Based Reasoning System

먼저 사람이 읽기 편하게 작성되어 있는 전자문서를 컴퓨터가 처리할 수 있는 형태의 자료구조로 표현하는 일이 선행되어야 한다. 이를 위해서 한국어 및 영어 단어를 먼저 구분하고, 각 언어에 적용할 수 있는 형태소 분석기 라이브러리를 사용하였다. 한국어의 경우에는 한국과학기술원에서 개발한 한나눔 형태소 분석기[4]를, 영어의 경우에는 루씬[5]을 이용하여 자연어로부터 단어를 추출해 내었다. 이렇게 추출된 단어 집합들을 이용하여 핵심어 비교 절차를 거친 후 최종 결과를 산출하게 된다.

4.1 순서도

사례 기반 사전판정 심사 시스템을 실제로 개발하기 위해서는 위에서 제안된 방법들을 실제

4) http://kldp.net/projects/hannanum

5) http://lucene.apache.org/

```
We have 24 classes

class score(alpha): 0.4086
cosine similarity(beta): 0.4652
final value: 0.4369

---the new document would be belong in the classes below---
the class A is a [해당] candidate with credibility 0.9, the number of keywords in this class matched 17 over 35
the class F is a [해당] candidate with credibility 0.5, the number of keywords in this class matched 10 over 35
the class K is a [비해당] candidate with credibility -0.75, the number of keywords in this class matched 8 over 35

---the new document would be similar with the documents below---
the document L is similar with credibility 1.00
the document S is similar with credibility 0.71
the document W is similar with credibility 0.62

Resultingly, the new document might be [해당] with 43.69% credibility

the CBR process is over, Thank you!
```

〈Figure 7〉 Output of Case-based Reasoning System

프로그래밍 절차로 표현되어야 한다. 따라서 다음 <Figure 6>에서 보는 것처럼, 프로그램의 시작과 처리 과정을 상세히 표현한 순서도를 제작하였다.

사례 기반 추론 시스템은 다음과 같은 순서로 작업을 진행한다.

1) 문서의 형태소 분석 과정
2) 신규 및 기존 사례 문서의 TF-IDF 결과 계산
3) 신규 문서와 기존 사례의 키워드 기반 유사도 비교 결과 (α)
4) 신규 문서와 계통의 유사도 비교 결과 (β)
5) 통합 결과 (γ)
6) 결과 출력

4.2 실행 결과

데모 프로그램을 실행하면 자동으로 다음 <Figure 7>과 같은 결과창이 나타난다.

이는 사례 기반 사전판정 결과와 그 추론 과정의 중간 결과를 모두 나타내고 있는데, 이 결과는 다음과 같이 해석할 수 있다.

데모 환경의 데이터베이스에는 총 24개의 계통이 등록되어 있고 문서-계통 유사도는 0.4086, 문서-문서의 코사인 유사도는 0.4652, 두 점수의 합산은 0.4367로 제시되었다.

신규문서와 계통간 유사도 계산 결과는 다음과 같다.

- 신규문서는 0.9 신뢰도로 [해당] 계통인 A계통과 유사한데, 총 35개 핵심어 중 17개의 매칭이 발생하였다.
- 신규문서는 0.5 신뢰도로 [해당] 계통인 F계통과 유사한데, 총 35개 핵심어 중 10개의 매칭이 발생하였다.
- 신규문서는 0.75 신뢰도로 [비해당] 계통인 K계통과 유사한데, 총 35개 핵심어 중 8개의 매칭이 발생하였다.

한편, 신규문서와 기존 문서간 유사도 계산 결과는 다음과 같다.

- 신규문서는 [해당] 판정이 내려져 있는 문서 L과 1.00 정도로 유사하다.
- 신규문서는 [해당] 판정이 내려져 있는 문서 S와 0.71 정도로 유사하다.
- 신규문서는 [비해당] 판정이 내려져 있는 문서W와 0.62 정도로 유사하다.

결과를 모두 종합해 보면 예시에서의 신규문서는 전략물자에 [해당] 한다고 볼 수 있으며, 이 결과의 신뢰도는 43.69%이다.

5. 결론 및 시사점

5.1 관련분야에의 기여

본 연구에서는 전략물자 수출통제 실무에 필요한 사례 기반 전문가 시스템을 원자력 분야에 맞추어 처음으로 설계하고 그 응용 가능성을 확인하였다. 이를 위하여 시스템 설계의 이론적 근거가 되는 Kim et al. (2014)에서 제안한 전략물자 계통의 핵심어 추출방법의 실험적 한계를 보완하고자 원자력 전문가를 통한 반자동식 핵심어 추출 실험을 추가적으로 진행하였다. 또한 원자력 계통과 문서 간 유사도 계산 방법을 새롭게 제안해 이를 전통적인 TF-IDF 유사도 점수와 통합하여 원자력 분야에 특화된 새로운 문서 분류 기준을 도출하고 전략물자 사전판정의 해당 또는 비해당 판정 근거로 활용하도록 하였다. 본 연구진은 실무적인 관점에서 이러한 과정을 아우르는 데모 프로그램을 제작하여 관련 전문가의 평가를 거쳤다.

이 시스템은 전략물자 관리에 있어 보다 효율적인 운영을 가능케 할 것이다. 또한 자동화된 시스템을 도입함으로써, 지식의 누출 및 관리의 위험 가능성을 줄일 수 있고, 전략물자의 불법 수출 가능성을 차단하며, 현장 심사관의 정확한 전략물자 통제를 지원하여, 국가 안전 보장을 넘어 세계 안전 보장에 큰 기여를 할 것으로 기대된다.

뿐만 아니라, 지금까지 다양한 분야에서 전문가 시스템 및 텍스트 마이닝에 대한 연구가 많이 있었지만, 이를 바탕으로 인공지능 개념을 적용한 전략물자라는 전문지식에 특화된 시스템 개발연구는 많지 않았다.

지능형 전략물자 수출통제 시스템을 개발하면서 축적된 경험은 전략물자뿐만 아니라 다른 통제 시스템과 관련된 분야의 의사결정을 위한 전문가 시스템을 구축하기 위한 기반 기술로서 활용될 수 있으며, 전문가 시스템의 연구 영역을 확장함으로, 새로운 인공지능 시스템 기술에 대한 연구를 촉진시킬 것이다.

5.2 향후 연구 과제

일반적으로 전문가 시스템은 컴퓨터가 특정 영역에서 전문가의 판단을 대신할 수 있도록 설계된 시스템 (Kendal and Creen, 2007)이다. 본 연구는 특히 원자력 물품 및 기술의 수출 심사에 특화된 사례 기반 추론을 활용하는 전문가 시스템을 제안하였다. 사례 기반 추론의 핵심인 사례의 유사도는 많은 방식을 통해 측정될 수 있는데, 수출 신청 문서는 자체의 텍스트뿐만 아니라 여러 가지 메타 정보를 포함하고 있기 때문에, 핵심어를 기반으로 한 유사도 검색뿐만 아니라 메타 정보를 활용한 유사도 검색 기능도 적용 가능할 것으로 보인다. 향후 연구에서는 메타 정보

를 활용하여 보다 다각적인 판별 기준을 확립하는 것이 필요하다.

또한, 지능형 수출 통제 시스템의 최종 목적인 사용자의 신속한 의사 결정을 돕기 위해서는, 꾸준한 피드백을 통해 사용자의 의사 결정 과정을 더욱 잘 반영하도록 사례 기반 (Case-base)의 기능이 점차 고도화되는 과정이 필요하다. 이를 위해서는 지식 기반 시스템에서 중요한 적응 학습 알고리즘을 적용하여, 사용자가 시스템을 사용할수록 시스템의 성능이 높아지도록 만들 필요가 있다.

보안성이 강한 원자력 수출 통제 분야의 업무 특성상 나타난 본 연구의 한계점으로는 연구기간 중 충분한 양의 문서 데이터를 확보하는 데에 어려움이 있었기 때문에, 핵심어 기반 유사도 비교 및 계통 기반 유사도 결과를 포함하는 새로운 프레임워크가 실무 데이터를 통하여는 검증되지 않았다는 점이다. 연구진은 대신 현장 전문가의 주관적 의견을 반영하여 데모 프로그램의 성능을 평가하였으며, 다량의 실제 데이터를 통한 검증 및 개선 또한 향후 반영되어야 할 과제이다.

5.3 기대효과

국제원자력기구 (IAEA)에 따르면 2030년 세계 원자력 발전소 수는 총 700기 이상으로 늘어나고, 그 시장 규모만 1200조원에 이를 것이라고 한다.[6] 현재 우리나라는 계속하여 원전 건설 사업을 추진하고 있으며, 기하급수적으로 늘어나는 사전 판정을 적시에 정확하게 수행하기 위해서는 기존 심사 시스템보다 월등히 성능이 향상된 심사 시스템이 필요하다. 심사관이 일일이 검토하고 대조하던 기존 전략물자 심사 시스템에 지능적인 수출 통제 전문가 시스템을 적용함으로써, 시간과 인력이 많이 소모되던 기존의 심사 시스템의 효율성과 효과성을 크게 향상 시킬 것으로 예상된다. 특히 전략물자의 불법 수출 가능성을 체계적으로 차단하고, 수출 품목의 적기 수출을 지원함으로써, 전략물자 수출 사업 증대에 큰 기여하게 될 것이다.

아직 세계적으로 원자력이라는 특정 전문지식을 활용한 지능형 전략물자 전문가 시스템은 존재하지 않으며, 수출 통제에 필요한 전략물자 기술 문서에 특화된 텍스트 마이닝 기법에 대한 연구는 부재하다. 원자력이라는 특정 전문지식을 다양한 분야에서 적용되고 있는 인공지능 기술 및 텍스트 마이닝을 이용함으로써, 독창적 전략물자 수출 통제 시스템을 개발하여, 이를 통해 전략물자 심사 시스템에 대한 국내 기술 및 경쟁력을 확보할 수 있으며, 전략물자 심사 자동화 시스템 관련 특허를 선점함으로써 국가적인 경쟁력을 높이고 이를 통하여 상당한 로열티 수입을 기대할 수 있다.

국제사회의 전략물자 수출 통제는 현재 매우 중요한 이슈이다. 우리나라 원자력 수출이 점차 증가함에 따라 국제사회가 주시하고 있는 상황에서, 지능형 전략물자 수출 통제 시스템은 국제사회 안에서 원자력 기술과 함께 선도해 나가는 기술로서 우리나라 수출 사업 증대에 기여할 것이다.

본 연구에서 개발되고 있는 지능형 전략물자 수출 통제 시스템은 수십만 건의 사전 판정 심사에 대해 심사관이 신속하고 객관적인 결론을 도출할 수 있게 도와줄 것이다. 이러한 효율적인 시스템을 통한 시간과 비용의 감소로 행정적인 부담을 줄일 뿐만 아니라, 경제적인 성과도 기대할 수 있다. 또한 전문가를 양성하기 위한 오랜

6) IAEA, http://www.iaea.org

시간 동안의 금전적, 물리적 투자를 대폭적으로 절약할 수 있게 되며, 부가가치가 높은 업무에 적용하여, 추가적인 경제적 효과를 가져 올 수 있을 것으로 기대된다.

참고문헌(References)

Aizawa, A., "An information-theoretic perspective of tf-idf measures," *Information Processing and Management*, Vol.39, No.1(2003), 45～65.

Al-Mubaid, H. and R. K. Singh, "A text-mining technique for extracting gene-disease associations from the biomedical literature," *International Journal of Bioinformatics Research and Applications*, Vol.6, No.3(2010), 270～286.

Ananiadou, S., T. Ohta, and M. K. Rutter, "Text Mining Supporting Search for Knowledge Discovery in Diabetes," *Current Cardiovascular Risk Reports*, Vol.7, No.1(2013), 1～8.

Ananiadou, S., B. Rea, N. Okazaki, R. Procter, and J. Thomas, "Supporting Systematic Reviews Using Text Mining," *Social Science Computer Review*, Vol.27, No.4(2009), 509～523.

Cao, Q., W. Duan, and Q. Gan, "Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach," *Decision Support Systems*, Vol.50, No.2(2011), 511～521.

Chen, Y. L., Y. H. Liu, and W. L. Ho, "A text mining approach to assist the general public in the retrieval of legal documents," *Journal of American Medical Informatics Association*, Vol.64, No.2(2013), 280～290.

Corley, C. D., D. J. Cook, A. R. Mikler, and K. P. Singh, "Text and Structural Data Mining of Influenza Mentions in Web and Social Media," *International Journal of Environmental Research and Public Health*, Vol.7, No.2 (2010), 596～615.

Feldman, R. and J. Sanger, *The text mining handbook: advanced approaches in analyzing unstructured data*, Cambridge University Press, Cambridge, 2007.

Firdhous, M., "Automating Legal Research through Data Mining," *International Journal of Advanced Computer Science and Applications*, Vol.1, No.6(2012), 9～16.

Ghose, A., "Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics," *IEEE Transactions on Knowledge and Data Engineering*, Vol.23, No.10(2011), 1498～1512.

Gupta, V. and G. S. Lehal, "A Survey of Text Mining Techniques and Applications," *Journal of Emerging Technologies in Web Intelligence*, Vol.1, No.1(2009), 60～76.

Hu, X., J. S. Downie, and A. F. Ehmann, "Lyric Text Mining in Music Mood Classification," *Proceedings of the 10th International Society for Music Information Retrieval Conference*, (2009), 411～416.

Hulth, A., "Improved Automatic Keyword Extraction Given More Lin-guistic Knowledge," *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, (2003), 216～223.

Hung, J. I., "Trends of e-learning research from 2000 to 2008: Use of text mining and bibliometrics," *British Journal of Educational Technology*, Vol.43, No.1(2012), 5～16.

Hur, J., A. D. Schuyler, D. J. States, and E. L. Feldman, "SciMiner: web-based literature mining tool for target identification and functional enrichment analysis," *Bioinformatics*, Vol.25, No.6(2009), 838～840.

Jessop, D. M., S. E. Adams, E. L. Willighagen, L. Hawizy, and P. Murray-Rust, "OSCAR4: a flexible architecture for chemical text-mining," *Journal of Cheminformatics*, Vol.3, No.1(2011), 41～52.

Kendal, S. L. and M. Creen, *An introduction to knowledge engineering,* Springer London, London, 2007.

Kim, U., H. Kim, M. Y. Yi, and D. Shin, "Nuclear exports control system using semi-automatic keyword extraction," *International Journal of Information and Electronics Engineering*, Vol.4, No.4(2014), 293～297.

Kodratoff, Y., "Knowledge discovery in texts: a definition, and applications." *Foundations of Intelligent Systems, Proceedings of the 11th International Symposium,* (1999), 16～29.

Kozomara, A. and S. Griffiths-Jones, "miRBase: integrating microRNA annotation and deep-sequencing data," *Nucleic Acids Research*, Vol.39, No.1(2011), 152～157.

Krallinger, M., F. Leitner, and A. Valencia, "Analysis of Biological Processes and Diseases Using Text Mining Approaches," *Bioinformatics Methods in Clinical Research*, Vol.593, No.1(2010), 341～382.

Krallinger, M., A. M. Rojas, and A. Valencia, "Creating Reference Datasets for Systems Biology Applications Using Text Mining," *Annals of the New York Academy of Sciences*, Vol.1158, No.1(2009), 14～28.

Landeghem, S. V., F. Ginter, Y. V. D. Peer, and T. Salakoski, "EVEX: a pubmed-scale resource for homology-based generalization of text mining predictions," *Proceedings of the 2011 Workshop on Biomedical Natural Language Processing*, (2011), 28～37.

Lee, H. S., H. G. Song, and H. S. Lee, "Classification of Photovoltaic Research Papers by Using Text-Mining Techniques," *Applied Mechanics and Materials*, Vol.284, No.1 (2013), 3362～3369.

Lee, J., *Expert systems, principles and development,* bubyoungsa, Seoul, 1996.

Li, N. and D. D. Wu, "Using text mining and sentiment analysis for online forums hotspot detection and forecast," *Decision Support Systems*, Vol.48, No.2(2010), 354～368.

Liao, S., "Expert System methodologies and applications – a decade review from 1995 to 2004," *Expert Systems with Application*, Vol. 28, No.1(2005), 93～103.

Lin, F. R., L. S. Hsieh, and F. T. Chuang, "Discovering genres of online discussion threads via text mining," *Computers and Education*, Vol.52, No.2(2009), 541～495.

Liritano, S. and M. Ruffolo, "Managing the Knowledge Contained in Electronic Documents: a Clustering Method for Text Mining," *Proceedings of the 12th International Workshop on Database and Expert Systems Applications*, (2001), 454～458.

Liu, X., S. Yu, F. Janssens, W. Glanzel, Y. Moreau, and B. D. Moor, "Weighted hybrid clustering by combining text mining and bibliometrics on a large-scale journal database," *Journal of the American Society for Information Science and Technology*, Vol.61, No.6(2010), 1105～1119.

Macskassy, S. A., "Contextual linking behavior of bloggers: leveraging text mining to enable topic-based analysis," *Social Network Analysis and Mining*, Vol.1, No.4(2011), 355～375.

Navathe, S. B., and R. Elmasri, *Fundamentals of database systems*, Pearson Education, Upper Saddle River, NJ, 2000.

Netzer, O., R. Feldman,, J. Goldenberg, and M. Fresko, "Mine Your Own Business: Market-Structure Surveillance Through Text Mining," *Marketing Science*, Vol.31, No.3 (2012), 521～543.

Powers, D. M. W., "Evaluation: From precision, recall and f-measure to roc., informedness, markedness and correlation," *Journal of Machine Learning Technologies*, Vol.2, No.1 (2011), 37～63.

Prentzas, J. and I. Hatzilygeroudis, "Categorizing approaches combining rule-based and case-based reasoning," *Expert Systems*, Vol.24, No.2(2007), 97～122.

Rajpathak, D., R. Chougule,, and P. Bandyopadhyay, "A domain-specific decision support system for knowledge discovery using association and text mining," *Knowledge and Information Systems*, Vol.31, No.3(2012), 40 5～432.

Rak, R., A. Rowley, W. Black, and S. Ananiadou, "Argo: an integrative, interactive, text mining-based workbench supporting curation," *The journal of biological databases and curation*, (2012).

Vellay, S. G. P., L. N. E. Miller,, and G. Paillard, "Interactive Text Mining with Pipeline Pilot: A Bibliographic Web-Based Tool for PubMed," *Infectious Disorders - Drug Targets (Formerly Current Drug Targets - Infectious Disorders)*, Vol.9, No.3(2009), 366～374.

Wyner, A., R. Mochales-Palau, M.-F. Moens, and D. Milward, "Approaches to Text Mining Arguments from Legal Cases," *Semantic Processing of Legal Texts*, *Lecture Notes in Computer Science*, Vol.6036(2010), 60～79.

Xie, B., Q. Ding, H. Han, and D. Wu, "miRCancer: a microRNA-cancer association database constructed by text mining on literature," *Bioinformatics*, Vol.29, No.5(2013), 638～644.

Yan, X. W., Y. F. Zheng, C. Yuan, and M. Q. Duan, "Research of Expert System in Nuclear Power Plant," *Applied Mechanics and Materials*, Vol.409-410(2013), 1569～1572.

Yang, Y., "An evaluation of statistical approaches to text categorization," *Information retrieval*, Vol.1, No.(1-2)(1999), 69～90.

Yang. H., I. Spasic, J. A. Keane, and G. Nenadic, "A Text Mining Approach to the Prediction of Disease Status from Clinical Discharge Summaries," *Journal of American Medical Informatics Association*, Vol.16, No.4(2009), 596～600.

Yoon, J., "Detecting weak signals for long-term business opportunities using text mining of Web news," *Expert Systems with Applications*, Vol.39, No.16(2012), 12543～12550.

Abstract

Export Control System based on Case Based Reasoning: Design and Evaluation

Woneui Hong* · Uihyun Kim* · Sinhee Cho* · Sansung Kim* · Mun Yong Yi** · Donghoon Shin***

As the demand of nuclear power plant equipment is continuously growing worldwide, the importance of handling nuclear strategic materials is also increasing. While the number of cases submitted for the exports of nuclear-power commodity and technology is dramatically increasing, preadjudication (or prescreening to be simple) of strategic materials has been done so far by experts of a long-time experience and extensive field knowledge. However, there is severe shortage of experts in this domain, not to mention that it takes a long time to develop an expert. Because human experts must manually evaluate all the documents submitted for export permission, the current practice of nuclear material export is neither time-efficient nor cost-effective. Toward alleviating the problem of relying on costly human experts only, our research proposes a new system designed to help field experts make their decisions more effectively and efficiently. The proposed system is built upon case-based reasoning, which in essence extracts key features from the existing cases, compares the features with the features of a new case, and derives a solution for the new case by referencing similar cases and their solutions. Our research proposes a framework of case-based reasoning system, designs a case-based reasoning system for the control of nuclear material exports, and evaluates the performance of alternative keyword extraction methods (full automatic, full manual, and semi-automatic). A keyword extraction method is an essential component of the case-based reasoning system as it is used to extract key features of the cases. The full automatic method was conducted using TF-IDF, which is a widely used de facto standard method for representative keyword extraction in text mining. TF (Term Frequency) is based on the frequency count of the term within a document, showing how important the term is within a document while IDF (Inverted Document Frequency) is based on the infrequency of the term within a document set, showing how uniquely the term represents the document. The results show that the semi-automatic approach, which is based on the collaboration of machine and human, is the most effective

* Department of Knowledge Service Engineering, KAIST
** Corresponding Author: Mun Yong Yi
Department of Knowledge Service Engineering, KAIST
291 Daehak-ro, Yuseong-gu, Daejeon 305-70185 Hoegi-ro, Dongdaemun-gu, Seoul 130-722, Korea
Tel: +82-42-350-1613, Fax: +82-42-350-1610, E-mail: munyi@kaist.ac.kr
*** Korea Institute of Nuclear Nonproliferation and Control

solution regardless of whether the human is a field expert or a student who majors in nuclear engineering. Moreover, we propose a new approach of computing nuclear document similarity along with a new framework of document analysis. The proposed algorithm of nuclear document similarity considers both document-to-document similarity (α) and document-to-nuclear system similarity (β), in order to derive the final score (γ) for the decision of whether the presented case is of strategic material or not. The final score (γ) represents a document similarity between the past cases and the new case. The score is induced by not only exploiting conventional TF-IDF, but utilizing a nuclear system similarity score, which takes the context of nuclear system domain into account. Finally, the system retrieves top-3 documents stored in the case base that are considered as the most similar cases with regard to the new case, and provides them with the degree of credibility. With this final score and the credibility score, it becomes easier for a user to see which documents in the case base are more worthy of looking up so that the user can make a proper decision with relatively lower cost. The evaluation of the system has been conducted by developing a prototype and testing with field data. The system workflows and outcomes have been verified by the field experts. This research is expected to contribute the growth of knowledge service industry by proposing a new system that can effectively reduce the burden of relying on costly human experts for the export control of nuclear materials and that can be considered as a meaningful example of knowledge service application.

Key Words : Expert System, Export Control System, Nuclear Nonproliferation and Control, Cased Based Reasoning

Received: June 29, 2014 Revised: August 3, 2014 Accepted: August 24, 2014

저 자 소 개

홍 원 의
성균관대학교에서 컴퓨터공학 및 수학 학사학위를 취득하였으며 현재 KAIST 지식서비스공학과에 석사과정으로 재학 중이다. 연구 관심분야는 Knowledge Engineering, E-Learning, Cognitive Engineering 등이다.

김 의 현
홍익대학교에서 전자공학 학사학위를 취득하였으며 KAIST 지식서비스공학과에서 석사학위를 취득하였다. 현재 Tibero에서 빅데이터 분석 업무를 맡고 있다. 연구 관심분야는 Intelligent Systems, HCI, Big Data 등이다.

조 신 희
KAIST 경영과학과 학사학위를 취득하였으며 현재 KAIST 지식서비스공학과에 재학 중이다. 삼성SDS의 Business Intelligence 컨설팅 그룹에서 인턴으로 일한 경력이 있으며, 연구 관심분야는 Management Information Systems, Information Retrieval 등이다.

김 산 성
한동대학교에서 컴퓨터공학 학사학위를 취득하였고, KAIST 지식서비스공학과에서 석사학위를 취득하였다. 현재 KBS 기술연구소에 재직하고 있다. 연구 관심분야는 Big Data Analysis, Information Retrieval, HCI 등이다.

이 문 용
미국 Maryland 대학에서 정보시스템으로 박사학위를 취득하였다. 현재 KAIST 지식서비스공학과 교수, 학과장으로 재직중이며, IJHCS의 부편집장, AIS-THCI의 시니어 편집장을 맡고 있다. 연구 관심분야는 Knowledge Engineering, Business Intelligence, Semantic Web, HCI 등이다.

신 동 훈
가톨릭대학교에서 의학물리 석사학위를 취득하였고, 서울대학교 원자핵공학과에서 2007년 박사과정을 취득하였다. 현재 KINAC에서 선임연구원으로 재직하고 있다. 연구 관심분야는 Data Mining, Text Mining, Artificial Intelligence, Image Similarity, Nuclear Nonproliferation Policy and Implementation 등이다.