# 텍스트 분석을 활용한 국가 현안 대응 R&D 정보 패키징 방법론

현윤진* · 한희준** · 최희석*** · 박준형**** · 이규하***** · 곽기영****** · 김남규*******

# Methodology Using Text Analysis for Packaging R&D Information Services on Pending National Issues

Yoonjin Hyun* · Heejun Han** · Heeseok Choi*** · Junhyung Park**** ·
Kyuha Lee***** · Kee-Young Kwahk****** · Namgyu Kim*******

## Abstract

The recent rise in the unstructured data generated by social media has resulted in an increasing need to collect, store, search, analyze, and visualize it. These data cannot be managed effectively by using traditional data analysis methodologies because of their vast volume and unstructured nature. Therefore, many attempts are being made to analyze these unstructured data (e.g., text files and log files) by using commercial and noncommercial analytical tools. Especially, the attempt to discover meaningful knowledge by using text mining is being made in business and other areas such as politics, economics, and cultural studies. For instance, several studies have examined pending national issues by analyzing large volumes of texts on various social issues. However, it is difficult to create satisfactory information services that can identify R&D documents on specific national issues from among the various R&D resources. In other words, although users specify some words related to pending national issues as search keywords, they usually fail to retrieve the R&D information they are looking for. This is usually because of the discrepancy between the terms defining pending national issues and the corresponding terms used in R&D documents. We need a mediating logic to overcome this discrep'ancy so that we can identify and package appropriate R&D information on specific pending national issues. In this paper, we use association analysis and social network analysis to devise a mediator for bridging the gap between the keywords defining pending national issues and those used in R&D documents. Further, we propose a methodology for packaging R&D information services for pending national issues by using the devised mediator. Finally, in order to evaluate the practical applicability of the proposed methodology, we apply it to the NTIS(National Science & Technology Information Service) system, and summarize the results in the case study section.

Keywords : Association Rule Mining, Social Network Analysis, Text Mining, Topic Analysis

논문접수일 : 2013년 08월 30일 논문수정일 : 2013년 09월 19일 논문게재확정일 : 2013년 09월 20일

※ 위 연구결과는 미래창조과학부의 지원으로 KISTI에서 수행하는 국가과학기술지식정보서비스 구축 사업(N-13-NM-LU01-C01)의 결과임.

* 국민대학교 비즈니스IT전문대학원 석사과정, e-mail : yoonjin0630@naver.com
** 한국과학기술정보연구원 NTIS센터 선임연구원, e-mail : hhj@kisti.re.kr
*** 한국과학기술정보연구원 NTIS센터 NTIS사업팀장, e-mail : choihs@kisti.re.kr
**** 국민대학교 비즈니스IT전문대학원 석사과정, e-mail : aldoset@naver.com
***** 국민대학교 비즈니스IT전문대학원 석사과정, e-mail : lghzzang0@naver.com
****** 국민대학교 경영정보학부 교수, e-mail : kykwahk@kookmin.ac.kr
******* 국민대학교 경영정보학부 부교수, e-mail : ngkim@kookmin.ac.kr

## 1. 서 론

최근 모바일 기술을 포함한 ICT 기술의 발전으로 인해 정보 환경이 많은 변화를 겪고 있으며, 그에 따라 사용자의 규모뿐 아니라 이러한 사용자들에 의해 생성, 공유, 저장되는 데이터 양이 기하급수적으로 증가하고 있다. 이러한 현상은 데이터의 양 자체가 문제의 일부분이 되는 빅데이터(Big Data) 분석 기술에 대한 수요와 관심을 증가시키고 있다(O'Reilly Radar Team, 2011). 빅데이터는 기존의 전통적인 방법이나 도구로는 수집, 저장, 검색, 분석, 시각화가 어려운 정형 또는 비정형의 대규모 데이터를 의미하며[McKinsey Global Institute, 2011], 빅데이터 관련 기술은 향후 2~5년 내에 IT분야에서 자리잡을 주요 기술로 예상되고 있다[Gartner, 2012]. 또한 IDC(2012)는 세계 빅데이터 시장이 2010년 32억 달러에서 2015년에는 169억 달러 규모에 달할 것으로 전망하고 있으며, 2020년에는 전 세계 디지털 정보의 양이 2009년에 비해 44배 정도 증가할 것으로 전망하고 있다[IDC, 2011]. McKinsey Global Institute [2011] 역시 2018년까지 미국에서만 빅데이터 관련 분야에 14~19만 명의 전문 인력과 150만 명의 데이터관리 인력이 필요할 것으로 예측하였다. 국내에서도 새로운 세원 발굴을 통한 세수 증가, 의료/복지를 포함한 행정 전반의 효율성 제고, 교통량 최적화를 통한 교통혼잡비용 감소 등, 빅데이터 활용을 통해 최대 4.2조 원의 부가가치를 창출할 수 있다고 보고된 바 있다[이부형, 2012).

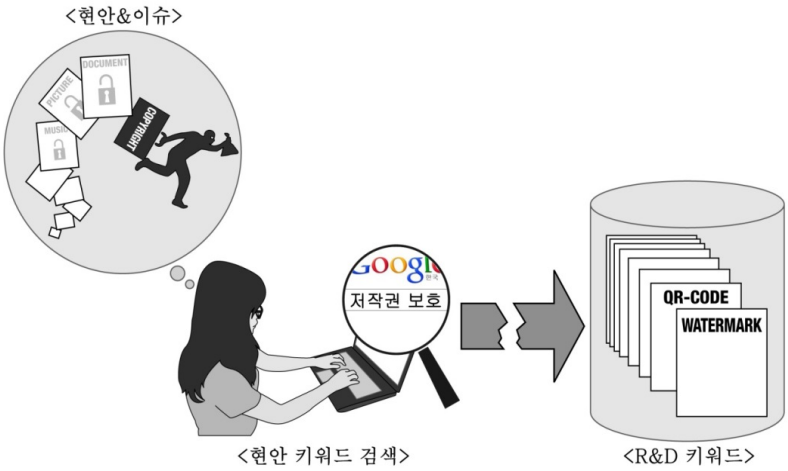
이처럼 빅데이터 기술에 대한 관심이 급증한 원인 중 하나는 스마트폰, 테블릿PC 등의 스마트 모바일 기기가 대중화됨에 따라 다양한 소셜 미디어를 통해 유통되는 비정형 데이터의 양이 급증한 것에서 찾을 수 있다. 특히 트위터(Twitter)와 페이스북(Facebook) 등을 통해 유통되는 텍스트 데이터는, 풍부한 정보나 의견을 거의 실시간으로 표현하고 있다는 특징으로 인해 연구자들의 많은 관심을 받고 있다. 이에 따라서 텍스트 형태의 비정형 빅데이터에 대한 분석이 텍스트 마이닝(Text Mining)이라는 이름으로 활발하게 이루어지고 있다. 텍스트 마이닝은 데이터 마이닝(Data Mining), 자연어 처리, 정보 검색, 전산 언어학, 토픽 추적(Topic Tracking) 등의 분야의 기술을 종합적으로 활용하여 대용량의 텍스트로부터 유용한 정보를 추출하는 과정(Hearst, 1999; Sebastiani, 2002)이라고 말할 수 있다. 기존에는 많은 기업들이 정형데이터를 활용하여 유용하고 잠재적인 정보를 얻기 위해 많은 노력을 기울였으나, 최근에는 텍스트 마이닝을 통해 소셜 미디어에서 매일 쏟아져 나오는 많은 양의 텍스트 형태의 비정형 데이터를 분석하여 좀 더 새롭고 유용한 정보를 얻어내고 있다. 대표적인 예로 국내 기업인 다음소프트는 소셜 미디어 분석 솔루션인 '소셜 메트릭스'라는 서비스를 통해 문맥 중심의 텍스트 마이닝 작업을 수행함으로써 각 데이터의 출현 원인에서부터 다른 데이터들과의 관계까지도 도식화하여 사용자에게 제공해 주고 있다[BLOTER.NET, 2012]. 이러한 분석의 대상이 되는 텍스트 데이터는 그 생성 속도가 매우 빠를 뿐 아니라 그 양이 실로 방대하고 정제되지 않은 형태로 유통되기 때문에, 정형의 소량 데이터를 대상으로 했던 전통적 분석 방식으로는 이들 데이터로부터 의미있는 정보를 추출하기가 매우 어렵다. 따라서 대량의 비정형 텍스트 데이터 속에 담긴 의미 있는 지식을 추출하고 요약해내기 위한 텍스트 마이닝 기술에 관심과 투자가 집중되는 것은 매우 자연스러운 현상이라 할 수 있다.

텍스트 마이닝을 통해 새로운 지식을 발굴하고자 하는 노력은 비즈니스 영역뿐 아니라 정치, 문화 등 다양한 영역에서 매우 활발하게 나타나고 있다. 특히 최근에는 정치・경제・사회문화 등 여러 현안 및 이슈들을 발굴하여 이를 의사결정

에 활용하고자 하는 시도가 활발하게 이루어지고 있다. 일례로, 2012년 대한민국의 주요 화두 중 하나였던 대통령선거에서 소셜 데이터의 분석을 통해 여론을 파악하는 선거 관련 서비스가 다수 등장했던 점을 들 수 있다. 선거 관련 서비스 중 와이즈넛의 경우 형태소 분석・텍스트 마이닝・자연어처리 등의 기술을 이용해 트위터, 페이스북, 블로그, 카페 등에 올라온 대선 후보 관련 버즈를 분석하기 위한 '버즈인사이트바이럴 지수(BVI)'를 자체 개발하기도 하였다. 또한 대선 후보자별 호불호는 물론 인품, 자질, 지지도 등의 여론 성향과 일자별 변화 추이를 비롯하여 버즈의 전파 경로 등 대선 후보 관련 정보를 '보드뷰' 형식으로 제공하는 서비스를 종합적으로 제공했다[etnews, 2012]. 와이즈넛 이외에도 실시간으로 기업, 제품, 정책 등에 대한 사용자의 반응과 이슈 등을 분석 보고하는 다이퀘스트의 '브람스(Brams)', 소셜 미디어에서 가장 화제가 되고 있는 사건, 인물 등을 분석해 제공해주는 코난테크놀로지(www.konantech.co.kr)의 '펄스K(www.pulsek.com)' 서비스, 다양한 소셜 빅 데이터를 수집 및 분석하여 제품, 사회 등과 관련된 이슈나 트렌드 등에 관한 정보를 제공하는 솔트룩스(www.saltlux.com)의 '트루 스토리(True Story)'서비스 등 다양한 형태의 소셜 데이터 분석 서비스가 제공되고 있다. 기존의 검색 서비스 전문 업체들의 경우, 언어분석 기반의 소셜 데이터 분석 서비스를 제공하고 있다. 즉, 검색포털사이트의 검색어 통계, 소셜 미디어의 게시글 및 댓글, 웹 페이지 상에서 이루어지는 사용자의 활동기록 등을 분석함으로써, 현재 이슈를 진단하고 더 나아가 미래를 예측하기 위한 기술에 대한 수요가 점점 증가하고 있다[류법모 외, 2012].

이처럼 빅데이터 분석을 통해 국가현안이나 이슈를 발굴하고자 하는 시도가 꾸준히 이루어져 왔음에도 불구하고, 빅데이터 분석을 활용하여 국가현안 및 이슈로부터 이와 관련된 R&D 문서를 효율적으로 제공하는, 즉 국가현안과 R&D 정보를 효과적인 방법으로 패키징하는 방안은 현재까지도 충분히 마련되지 않고 있다. 많은 사용자들은 특정 시기에 이슈가 되는 현안에 대해 여러 매체 또는 보도자료를 통해서 자세한 정보를 얻곤 한다. 일부 사용자의 경우 특정 현안에 대한 이와 같은 일반적인 정보뿐 아니라, 현안과 관련된 보다 전문적이고 체계적인 정보를 얻기 위해 기술문서, R&D 문서 등의 전문 자료를 검색하고자 한다. 하지만 만약 사용자가 현안 키워드에는 익숙하지만 관련 R&D 키워드에는 익숙하지 않다면, 사용자 본인이 직접 정확한 R&D 키워드를 입력하여 적절한 자료를 획득하는 경우는 극히 드물다. 이와 같은 현상이 발생하는 원인은, 사용자가 인식한 현안에 대한 키워드와 실제 R&D 분야에서 이와 관련하여 사용되는 키워드 간에 이질성이 존재한다는 측면에서 찾을 수 있다. 이러한 현상은 <그림 1>을 통해 보다 자세히 설명된다.

<그림 1>은 사용자가 저작권 침해 사례를 접하고, 이에 대해 보다 자세한 정보를 얻기 위해 '저작권 보호'라는 현안 키워드를 사용하여 자료 검색을 시도하는 예를 보여주고 있다. 즉 사용자가 '워터마크(Watermark)', 'QR-Code' 등 R&D 키워드에 익숙하지 않은 경우, 사용자는 원하는 기술 정보를 얻기 위해 '저작권 보호' 등의 익숙한 현안 키워드를 초기 검색어로 선택하는 경우가 대부분이다. 이처럼 사용자가 현안 키워드 기반의 검색을 시도하는 경우, 원하는 R&D 정보를 찾기 위해 수많은 시행착오를 겪어야 할 뿐 아니라, 이렇게 수많은 검색 시도를 통해 사용자가 원하는 R&D 문서를 항상 획득할 수 있는 것도 아니다. 이러한 한계는 사용자가 떠올리는 현안 키워드의 풀(Pool)과 R&D 문서의 키워드로 사용되는 풀 사이의 이질성으로 인해 나타난다. 따라서 현안 및 R&D 키워드의 이질성을 극복하기 위한




〈그림 1〉 현안 키워드와 R&D 키워드 간 이질성으로 인한 검색의 어려움

중간 장치가 필요하며, 이 중간 장치를 통해 각 현안 키워드와 각 R&D 키워드 간에 적절한 대응이 이루어져야 한다. 예를 들어, '저작권 보호'라는 현안 키워드를 '워터마크', 'QR-Code'라는 R&D 키워드에 대응시킴으로써, 사용자가 현안 키워드를 입력하였을 때 이에 대응되는 적절한 R&D 문서를 제공할 수 있게 된다.

이를 위해 본 연구에서는 현안 대응 R&D 패키징 방법론을 제안하고자 한다. 다음 장인 제 2장에서는 텍스트 마이닝, 연관관계 분석, 소셜 네트워크분석, 국가 R&D 정보 서비스 현황에 대한 관련 연구들을 간략하게 소개한다. 제 3장에서는 현안 대응 R&D 패키징 방법론을 좀 더 구체적으로 제시한다. 즉 현안 및 R&D 키워드 동시 출현 문서의 발굴, 현안 사전 및 R&D 용어집을 이용한 분석대상 용어 추출 과정, 추출된 키워드 간 연관규칙 도출, 연관규칙을 이용한 연관 키워드 네트워크 구축, 그리고 최단거리 및 최대경로 수에 기반한 현안 대응 R&D 매핑 등, 제안 방법론의 각 단계를 상세히 설명한다. 또한 제 4장에서는 사례 적용을 통해 제안방법론의 실제 활용 가능성을 평가하고, 마지막 장인 제 5장에서는 본 연구의 기여 및 한계 그리고 향후 연구방향을 제시한다.

## 2. 관련 연구

### 2.1 텍스트 마이닝

텍스트는 현실 세계에서 정보를 교환하고, 자신의 의사를 표현하는 방법으로 가장 널리 사용되는 수단이다[Witten, 2004]. 따라서 많은 연구자들이 풍부한 정보를 담고 있는 텍스트에 대한 분석을 통해 의미 있는 지식을 발견하기 위해 끊임없이 노력하고 있다. 텍스트 마이닝은 대용량의 방대한 텍스트로부터 구문 분석을 통해 유용한 정보를 추출하는 과정[Hearst, 1999; Sebastiani, 2002]이라고 말할 수 있다. 텍스트 마이닝의 활용 분야는 텍스트 형태로 된 정보를 사용하는 모든 분야를 아우를 정도로 매우 다양하다. 예를 들자면, 특정 기사(Article)의 원문(Source)를 파악하기 위한 연구[Metzler et al., 2005], 특정 범죄와 다른 범죄들 간의 유사성 측정을 통해 새로운 범죄를 발견하기 위한 연구[Fan et al., 2006], 텍스트 범주화(Categorization)를 통해 비구조적 저장소(Repository)를 구조화하기 위한 연구[Sebastiani, 2006] 등도 텍스트 마이닝의 범주에 포함된다. 특히

최근에는 다양한 성향을 가지는 사용자들의 소통 수단인 소셜 미디어를 통해 방대한 양의 텍스트 데이터가 공유됨에 따라, 소셜 미디어 데이터에 대한 텍스트 마이닝을 통해 기존의 데이터 분석에서는 찾을 수 없었던 새로운 유형의 지식을 찾기 위한 시도들이 활발하게 이루어지고 있다[(김인현, 2012; 최광선, 2012].

텍스트 마이닝은 데이터 마이닝, 자연어 처리, 정보 검색, 전산 언어학, 토픽 추적 등 여러 분야의 기술을 종합적으로 활용한다[Mooney and Bunescu, 2006; Rijsbergen, 1979]. 이러한 기술로 인해 기존의 데이터 마이닝 분야에서 해결해왔던 전통적 주제[김경재, 안현철, 2005; 허준, 김종우, 2008; 황인수, 2004]뿐 아니라, 더욱 폭넓은 주제에 대한 분석이 가능해졌다. 특히 자연어 처리 기술은 텍스트 마이닝의 성패를 좌우하는 핵심 기술이라고 할 수 있으며, 자연어 처리의 대상이 되는 텍스트는 분석 목적에 따라 행렬, 계층, 벡터 등의 다양한 형태로 표현된다[Stanvrianou et al., 2007]. SAS Enterprise Miner, IBM SPSS Modeler, R, Linguamatics 12E 등 대부분의 데이터마이닝 소프트웨어는 텍스트 마이닝 기능을 지원한다. 대부분의 사용 분석 도구에서 분석의 최소 단위는 각 문서가 되며, 여기서 문서란 제목, 요약, 본문, 문서전체 등 텍스트로 기술된 모든 데이터를 일컫는 폭넓은 개념으로 사용된다. 기본적으로 각 문서는 벡터 공간모델(Vector Space Model)[Albright, 2006; Salton et al., 1975]을 이용하여 표현되며, 각 문서에 사용된 용어(Term)의 빈도에 따라 해당 문서의 주제 및 특성이 요약된다. 대부분의 경우 용어의 단순 빈도수보다는 TF-IDF(Term Frequency-Inverse Document Frequency)[Han and Kamber, 2011]에 근거한 분석이 널리 활용된다. 이 개념은 어떤 문서 D에서 용어 A와 B가 동일한 빈도수로 발생하였을 때, A가 다른 문서들에서도 일반적으로 자주 발생하는 용어라면 문서 D에서 더 중요하게 사용되는 용어는 A가 아니라 B라는 인식을 전제로 한다. 빈도수에 기반한 분석에서 각 문서는 용어 수만큼의 차원을 갖게 되며, “(문서 수) × (용어 수)”로 표현된 행렬의 각 셀에 각 문서에서 해당 용어가 나타난 빈도수를 기재함으로써 모든 문서를 행렬화할 수 있다. 하지만 문서에 포함된 용어의 수는 일반적으로 매우 많기 때문에, 문서간 유사성 측정을 위해 각 문서는 SVD (Singular Value Decomposition) 등의 차원 축소 기법을 통해 저장된다[Albright, 2006]. 상용 텍스트 마이닝 도구는 이러한 이론을 기본으로 하여 파싱, 필터링, 클러스터링 등의 작업을 수행한다. 이러한 작업의 결과는 이 자체로도 문서 분류, 토픽 추출 등의 작업에 사용될 수 있을 뿐 아니라, 기존의 마이닝 분석 모델인 의사결정나무, 인공신경망 등 후속 분석의 입력으로 사용될 수도 있다.

## 2.2 연관관계 분석

데이터 마이닝은 방대한 데이터로부터 유용한 정보나 패턴을 추출하는 기법으로, 통계적 기법, 인공지능 기법 등을 통해 연관관계(Association) 분석, 분류, 군집화 등의 여러 가지 지식을 창출하는 과정[Han and Kamber, 2011]에 널리 적용되고 있다. 특히 연관관계 분석[Agrawal and Srikant, 1994]은 데이터들의 빈도수와 동시 발생 확률을 이용하여 데이터와 데이터 간의 관계를 찾고 이를 규칙으로 표현하는 분석 기법으로, 장바구니 분석, 인터넷 쇼핑몰 추천시스템, 교차판매, 매장 배치, 카탈로그 설계, 판촉전략 수립 등 다양한 분야[김남규, 2008; 안현철 외, 2006; 윤성준, 2005; 이연정, 김경재, 2013; Wang et al., 2007]에서 활용되고 있다.

연관관계 분석을 통해 도출된 규칙들을 평가하기 위해 다양한 흥미성 척도가 고안되어 왔으며, 그 중 신뢰도(Confidence), 지지도(Support), 향

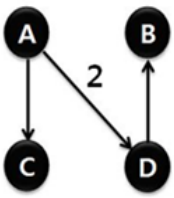
상도(Lift)의 세 가지 척도가 가장 널리 사용된다. 도출된 규칙의 향상도가 1 이상이면서 지지도와 신뢰도가 각각 최소지지도와 최소신뢰도 이상으로 나타날 때, 해당 규칙은 강한 연관성을 나타내는 것으로 간주된다[박우창 외, 2003]. 연관규칙 (A → B)에 대해, 지지도는 전체 트랜잭션 수 대비 A와 B가 동시에 출현한 트랜잭션 수의 비율을 의미하며, 신뢰도는 A를 포함하는 트랜잭션 중 A와 B를 함께 포함하는 트랜잭션의 비율을 의미한다. 즉, 특정 연관규칙에 대한 신뢰도는 조건에 맞는 결과가 얼마나 자주 적용될 수 있는지를 나타내고, 지지도는 연관규칙 자체가 얼마나 믿을 만한 것인지를 의미한다고 할 수 있다. 한편 향상도는 A와 B의 상관관계를 나타내는 척도로서, 그 값이 1이면 A와 B가 독립적인 관계를 나타내고, 1보다 큰 경우에는 양의 상관관계, 1보다 작은 경우에는 음의 상관관계를 나타낸다.

연관관계 분석은 그 자체로도 위와 같은 다양한 분야에 활용되어 왔지만, 최근 소셜 네트워크 분석(SNA : Social Network Analysis), 텍스트 마이닝 등의 분야의 비약적인 발전으로 인해 그 활용 범위가 더욱 다양해졌다. 즉 분석 대상 데이터에 대한 연관관계 분석을 실시하여 관심 항목간 연관성을 도출하고, 각 항목과 이들간 연관성을 네트워크로 도식화함으로써 보다 다층적인 분석을 실시할 수 있게 된 것이다[조인동, 김남규, 2011]. 또는 인터넷 문서, 뉴스 기사, 소셜 미디어 등에 대한 분석을 통해 연관어 맵을 도출하거나 문서간 분류를 수행하는 융용의 경우도, 기본적으로 연관관계 분석의 원리를 이용하는 것으로 파악될 수 있다. 본 연구에서 제안하는 방법론은 이러한 기존의 연구 성과를 통합 활용한 것으로, 텍스트 분석을 통해 용어간 동시출현 집합을 생성하고, 이에 대한 연관관계 분석과 소셜 네트워크 분석을 통해 국가 현안에 대응되는 R&D 정보를 식별하고자 한다.

### 2.3 소셜 네트워크분석

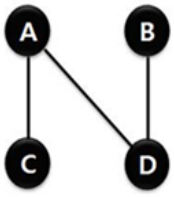
소셜 네트워크분석은 집단 내 개체의 연결 상태 및 연결 구조의 특성을 계량적으로 파악하여 시작적으로 표현하는 분석 기법으로[김용학, 2003], 유전 네트워크(Kauffman, 1993), 교통 네트워크, 조직 네트워크[최창현, 2006] 등의 구조 분석에도 널리 이용되고 있다. 개체(Node)와 개체 간의 관계(Link)로 구성되는 소셜 네트워크는, 그래프(Graph)를 통해 시각적으로 표현될 수도 있고 매트릭스의 형태로 수치적으로 표현될 수도 있다. 매트릭스는 소셜 네트워크 데이터를 표현하는 가장 기본적인 수단으로, 행(Column)과 열(Row)이 만나는 셀에 특정값을 표시함으로써 행과 열 사이의 관계를 표시하는 방법이다. 매트릭스에서 관계를 표현하는 기본적인 방법은, 두 행위자 간에 관계가 존재하면 '1', 그렇지 않으면 '0'으로 표현하는 것이다. 한 매트릭스에서 i행 j열에 위치한 셀의 경우 그 값 a는 a(i, j)라고 표현하고, 이러한 수치는 방향성과 연결 강도를 나타낼 수도 있다[손동원, 2002]. <그림 2(a)>와 같이 방향이 있는 네트워크의 경우, 이에 대응되는 매트릭스에서 각 행은 영향이나 정보를 주는 노드를 나타내고 각 열은 영향이나 정보를 받는 노드를 나타낸다. 관계의 방향성 없이 연관성만을 나타내는 <그림 2(b)>의 네트워크의 경우 행과 열이 나타내는 정보의 차이는 없다[강은영, 곽기영, 2011].

소셜 네트워크분석에서 네트워크 연결 구조의 특성을 파악하기 위한 대표적 측정 지표는 밀도(Density), 중심성(Centrality) 및 집중도(Centralization) 등이 있다[Freeman, 2008; Scott, 2000]. 특히 밀도는 네트워크에서 노드 간의 연결 정도 수준을 나타내며, 네트워크 내의 라인의 개수, 즉 연결 정도(Degree)로 측정할 수 있다. 하지만 연결 정도를 바탕으로 한 밀도와 같이 바로 인접한 노드와의 직접적 연결만을 분석하는 것은 네트워




| | A | B | C | D |
|---|---|---|---|---|
| A | - | 0 | 1 | 2 |
| B | 0 | - | 0 | 0 |
| C | 0 | 0 | - | 0 |
| D | 0 | 1 | 0 | - |

(a) 방향성 네트워크




| | A | B | C | D |
|---|---|---|---|---|
| A | - | 0 | 1 | 1 |
| B | 0 | - | 0 | 1 |
| C | 1 | 0 | - | 0 |
| D | 1 | 1 | 0 | - |

(b) 무방향성 네트워크

〈그림 2〉 네트워크의 매트릭스 표현 예

크 내 노드의 상대적 위치나 네트워크 자체를 이해하기 어렵다는 한계를 갖는다. 이러한 한계는 네트워크 내의 노드가 다른 노드들과 직접 연결되기보다는 대다수의 노드들과 간접적 연결관계를 갖는다는 특성에 기인한다. 따라서 네트워크 내의 노드들의 상대적 차이를 이해하기 위한 다양한 방법이 사용되고 있으며, 주로 노드 간의 거리(Distance), 즉 두 노드가 얼마나 가깝게 위치하고 있는지에 따라 두 노드의 밀접한 정도를 표현하는 방법이 사용된다. 노드들은 서로 직·간접적인 라인(Line)으로 연결되며, 그래프 상에서 이러한 링크의 전후 관계를 궤적(Walk)이라고 한다. 또한 노드와 라인으로 이루어지는 궤적을 경로(Path)라고 한다. 경로길이(Path length)는 경로에 포함된 라인의 수로 측정되며, 두 노드 간 복수 경로가 존재할 경우, 가장 짧은 경로를 최단경로(Geodesic, Geodesic Distance)라 한다. 일반적으로 두 노드간 거리는 두 노드간 최단경로의 길이를 의미한다.

짧은 거리로 연결된 노드들은 상호작용의 빈도와 강도가 높기 때문에, 그렇지 않은 노드들에 비해 더 강한 연결관계를 갖는 것으로 인식된다. 하지만 때로는 짧은 거리보다는 얼마나 다양한 경로를 통해 두 노드가 연결되고 있는지가 더욱 의미 있는 경우가 있다. 이는 두 노드가 서로에게 도달하기 위한 복수의 연결경로를 갖는 경우, 서로의 연결이 단절될 가능성이 작아서 보다 안정적이고 신뢰성 높은 관계를 유지하고 있다는 것을 의미하기 때문이다. 이 때, 두 노드 사이에 중복되지 않는 라인으로 구성된 연결경로를 플로우(Flow)라 하며, 플로우의 개수가 많을 수록 두 노드의 대체 경로 수가 증가하므로, 두 노드간의 연결관계가 강해진다. 또한 플로우는 라인 연결성(Line Connectivity)의 개념과 밀접한 관련을 갖는다. 즉, 두 노드간의 라인 연결성은 두 노드간 연결경로를 없애기 위해 제거되어야 하는 최소 라인의 수, 즉 두 노드 간의 최대 플로우(Maximum Flow)를 의미한다. 이는 어떤 정보를 접함에 있어, 그 정보가 얼마나 빨리 도착했는지 보다는 그 정보를 얼마나 많이 반복적으로 들었는지가 정보의 신뢰성을 결정하는데 더 중요한 역할을 한다는 측면으로 이해될 수 있다[곽기영, 2013].

소셜 네트워크분석을 위해 UCINET, NetMiner, R, 그리고 NodeXL 등 유용한 분석 툴이 제공되고 있을 뿐 아니라, 전통적인 네트워크 이론에 기반한 다양한 중심성 척도들이 편리하게 계산되기 때문에 점차 그 활용 분야가 확대될 것으로 기대된다. 즉 이 기법을 활용하고자 하는 연구자는 노드를 식별하고 노드들 간의 연결 라인을 정의하는 과정을 통해 네트워크를 형성함으로써, 그 이후의 분석 단계는 다양한 척도들을 계산해 주는 분석 툴의 도움으로 비교적 쉽게 수행할 수 있다. 하지만 사회의 모든 현상이 노드와 라인으로 직접 모델링 되는 것은 아니다. 예를 들어 본 연구에서 다루고 있는 현안 및 R&D 키워드간의 연관 네트워크의 경우, 키워드를 노드로 간주할 때 이들 간의 라인은 어떤 의미를 가져야 하는지에 대

한 판단이 이루어져야 한다. 본 연구에서는 키워드간의 연관성, 즉 해당 키워드들이 단위 논문에서 동시에 출현하는 정도를 노드간 라인으로 사용하고자 하며, 연관성을 나타내는 신뢰도 및 지지도의 도출을 위해 연관규칙 마이닝 기법을 사용하고자 한다. 또한 연관규칙을 네트워크 입력 데이터로 사용하기 위해, 다양한 형태의 네트워크 데이터를 생성하는데 있어 가장 효과적이면서도 유연한 방법인 DL(Data Language)을 이용하고자 한다. DL은 풀매트릭스(Fullmatrix), 노드리스트(Nodelist), 엣지리스트(Edgelist)의 총 3가지 형식으로 표현되며, 특히 본 연구에서는 그 중에서도 엣지리스트를 활용하여 네트워크 데이터를 구축하고자 한다.

### 2.4 국가 R&D 정보 서비스 현황

정부차원에서 진행하는 국가연구개발사업에 대한 관리 및 정보의 관리・유통체계의 구축은 [과학기술기본법]의 제12조, 제20조, 제26조 및 동법 시행령에 함께 규정되어 있다[류범종, 2003]. 초기 단계에는 국가연구개발사업의 성과물에 대한 관리가 효율적으로 이루어지지 않아, 정부의 막대한 투자에 비해 실제 성과는 매우 미흡했던 것으로 알려져 있다. 그 후, 2005년 '국가연구개발사업 등의 성과평가 및 성과관리에 관한 법률'이 제정되고 이를 수행하기 위한 '연구성과관리활용기본계획(안)'이 수립되면서 국가연구개발사업 성과관리 및 활용과 함께 이에 대한 체계적인 관리가 본격화되기 시작했다[김문수 외, 2008]. 이에 따라 국가연구개발사업의 효율적인 기획 및 관리 체제를 구축하여 국가 연구개발사업의 활동 주체간 원활한 정보 유통 체제와 효율적이고 효과적인 연구개발 정보 지원체제를 구축하는 등, 성과 통합관리에 대한 연구[류범종, 최기석, 2004; 신성호 외, 2011]들이 활발히 이루어졌다.

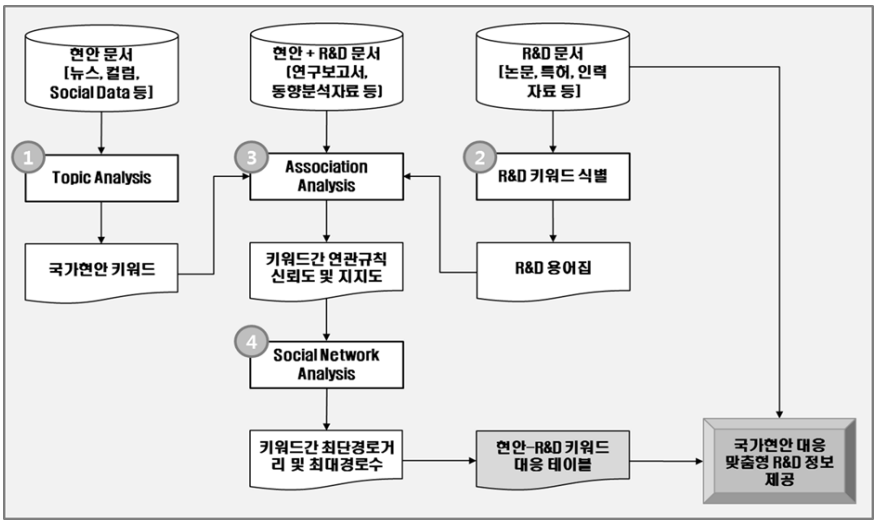
이처럼 국가 R&D 정보에 대한 관리 및 유통체계가 의미 있는 성과를 거둠에 따라, 국가 R&D 정보의 통합 관리뿐 아니라, 국가 R&D 정보를 활용하여 사용자에 효율적으로 서비스하는 방안에 대한 연구가 관심을 받고 있다. 대표적 연구로는 전 부처의 모든 국가 R&D 정보에 식별체계를 부여해 과제 관련 성과, 참여 인력 정보, 관련 장비, 성과와 관련된 인력정보 등 정보간 관련 연계정보를 바로 링크해서 볼 수 있는 참조연계서비스[권이남, 김재수, 2007], 국가 R&D 참여인력 데이터베이스를 활용하여 네트워크가 협소한 개인연구자나 중소기업이 정부 출연연구소나 대학의 전문가들로부터 기술자문을 받을 수 있는 의사소통 채널을 제공함으로써 네트워크를 연결시켜주는 전문가 검색대행서비스[양명석 외, 2010] 등을 들 수 있다.

이렇게 국가 R&D 정보를 통합 관리하고, 그 정보를 활용하여 효과적으로 서비스하기 위한 방안에 대한 연구가 활발하게 수행되었지만, 주어진 국가 현안에 대응하는 R&D 정보를 수월하게 제공하기 위한 현실적인 패키징 방안에 대한 연구는 상대적으로 매우 부족하다. 즉 R&D 정보는 체계적으로 구조화되어 저장되어 있지만, 이러한 정보를 사용자들이 궁금해하는 현안과 연결시켜 제공하기 위한 중간 장치가 부족했다고 할 수 있다. 따라서 이러한 고급 R&D 정보가 충분히 활용되기 위해서는, 사용자들이 떠올리는 현안 키워드와 실제 R&D 자료에 사용된 R&D 키워드를 적절하게 매핑시키기 위한 방안에 대한 연구가 반드시 필요하다.

## 3. 현안 대응 R&D 정보 패키징 방법론

### 3.1 연구 범위 및 용어 정의

본 절에서는 현안 대응 R&D 정보 패키징을 위한 방법론을 제시하고자 한다. 여기서 '현안'이란 정책입안자가 정한 정책을 비롯하여 각종 연구소의 정책보고서 및 언론, 소셜 미디어 등에서 거론




〈그림 3〉 전체 연구 개요

되는 정치, 경제, 사회문화적 이슈까지 포함한 광범위한 내용을 의미한다. '현안 키워드'란 하나의 현안 주제를 구성하고 있는 핵심 용어들을 의미한다. 또한 'R&D(Research and Development)'정보는 국가연구기관 및 민간연구소에서 이루어지는 모든 연구개발활동과 관련된 자료를 의미하며, 'R&D 키워드'란 좁게는 R&D 자료에 명시된 핵심어 항목을 의미한다.

본 연구의 기본적인 흐름은 각종 시장정보 및 R&D 자료에서 추출한 국가현안 주제 관련 키워드 및 R&D 관련 키워드 집합을 최소 분석 단위로 정의하고, 이 집합에 대한 연관관계 분석을 수행하여 연관규칙을 도출하는 것이다. 이렇게 도출된 연관규칙을 기반으로 연관 키워드 네트워크를 구축하고, 최단거리 및 최대경로 수에 기반한 현안과 R&D 대응 키워드 테이블을 제공하고자 한다. 전체 연구 개요는 <그림 3>에 제시되어 있다. 단 텍스트 분석의 품질 향상을 위해 필요한 표준화 작업, 동음이의어, 이음동의어, 그리고 유사어를 정제하기 위한 일련의 작업은 본 연구의 범위에서 제외하였다.

<그림 3>에서 주요 분석이 수행되는 과정은 (1)~(4)의 직사각형에 표시되어 있다. 우선 (1)과 (2)는 현안문서 및 R&D 문서로부터 토픽 분석을 통해 현안 키워드와 R&D 키워드를 추출하는 과정을 나타낸다. 구체적으로 분석대상 용어를 추출하기 위해 활용되는 현안 사전은 정책보고서, 뉴스, 소셜 데이터 등의 국가현안 관련 문서들에 대한 토픽 분석을 통해 추출된 키워드의 집합을 목록화하여 구축한다. 또한 R&D 용어집은 전문학술지 또는 과학기술 특허 등의 자료에 명시된 R&D 키워드를 식별하여 R&D 키워드 집합을 추출한 후 목록화하여 구축한다. (1)과 (2) 부분은 일반적인 토픽 분석의 절차를 따르므로 본 논문에서 자세히 다루지는 않는다. 토픽 분석은 기본적으로 각 문서에 출현한 용어들의 TF-IDF 값에 근거하여 이루어지며, 문서의 주제 형성에 기여도가 높은 어휘들의 집합으로 각 토픽이 구성된다. 하나의 토픽은 여러 어휘로 구성되며, 하나의 토픽에 대응되는 문서는 여러 개 존재할 수 있다. 또한 각 문서는 여러 토픽에 연관될 수 있다. 토픽 분석을 통한 결과의 예는 본 논문 제 4.2.1절에 나타나있다. 본 연구의 핵심 부분은 (3)과 (4)에 해당하는 부분으로, 현안 + R&D 문서(즉, 현안과 R&D 정보가

혼재되어 나타나는 문서)에 대한 분석을 통해 현안-R&D 키워드 대응 테이블을 도출하는 과정이다. 현안+R&D 문서에 대한 본격적인 분석에 들어가기 전에, 분석의 품질을 향상시키고 분석 시간을 단축시키기 위해 주요 용어를 제외한 다른 용어들을 문서에서 제거하는 정제 작업이 필요하다. 이 과정에서 (1)과 (2)에서 도출한 현안 사전과 R&D 용어집이 활용된다. 즉, 현안 사전과 R&D 용어집에 등록된 어휘의 합집합을 토픽 분석의 Start List로 적용시킴으로써, 관심 용어의 집합으로만 구성된 파싱 결과를 얻을 수 있다. 분석대상 용어 추출이 완료되면 이에 대한 연관관계 분석을 수행하여 현안 및 R&D 키워드 간 연관규칙을 도출할 수 있다. 즉 하나의 문서에 동시에 자주 언급되는 키워드들 간의 관계를 파악하고자 하며, 이 과정은 지지도와 신뢰도 등의 척도를 사용하여 수행한다. 본 연구에서 사용하고자 하는 대표적인 연관성 척도인 지지도(Sup.)와 신뢰도(Conf.)의 간략한 수정 정의는 다음과 같으며, 자세한 분석 과정은 제 3.2절에서 소개한다.

$Sup.(A \rightarrow B)$ = *(현안 키워드 A와 R&D 키워드 B가 동시에 명시된 문서의 수) / (전체 문서의 수)*

$Conf.(A \rightarrow B)$ = *(현안 키워드 A와 R&D 키워드 B가 동시에 명시된 문서의 수) / (현안 키워드 A가 명시된 문서의 수)*

이렇게 도출된 현안과 R&D 키워드 간의 연관규칙을 통해 각 현안에 대응되는 R&D 키워드를 식별할 수 있다. 하지만 이는 직접적인 연관성을 갖는 대응 관계만 식별한 것으로, 다른 매개 키워드를 통해 간접적이지만 강한 연관성을 갖는 대응 관계를 파악하지 못한다. 예를 들어(저작권 → 워터마크), (워터마크 → 인증)의 두 가지 규칙만 존재하는 경우, (저작권 → 인증)이 연관성을 가질 수 있는 가능성을 점검하지 못한다는 한계를 갖는다. 이러한 한계를 극복하고 보다 확장된 대응 관계를 파악하기 위해, 위에서 도출된 연관규칙을 소셜 네트워크로 구축하여 분석한다. 이 때, 네트워크의 간선으로는 방향성, 무방향성 간선을 모두 사용할 수 있는데, 신뢰도 기반 분석의 경우, 방향성 그래프가, 지지도 기반 분석의 경우 무방향성 그래프가 주로 활용된다. 또한 제안하는 방법론에서 소셜 네트워크 상의 각 노드가 현안 키워드를 의미하는지, R&D 키워드를 의미하는지, 또는 현안 키워드와 R&D 키워드를 동시에 의미하는지 파악하는 것은 매우 중요하다. 이는 궁극적인 결과로 각 현안에 대응되는 R&D 키워드를 식별하는 것이 연구의 목표이기 때문이다. 이를 위해 네트워크 상의 각 노드는 속성값으로 '현안', 'R&D', '현안+R&D'의 값을 갖게 된다. 이처럼 소셜 네트워크가 구축되면 네트워크 내의 노드 사이의 최단거리 및 최대경로 수를 도출하고, 이에 기반하여 현안-R&D 키워드 대응 테이블을 도출한다. 소셜 네트워크를 구축하고, 이에 대한 분석을 통해 현안-R&D 키워드 대응 테이블을 도출하는 과정은 제 3.3절에서 자세히 소개한다. 이러한 모든 과정을 통해, 사용자는 현안 키워드를 입력하여 관련 R&D 정보의 패키지를 쉽게 획득할 수 있는 국가 현안 대응 R&D 정보 서비스를 제공받을 수 있다.

## 3.2 현안 및 R&D 키워드 간 연관관계 분석

### 3.2.1 현안 및 R&D 키워드 포함 자료 선정

현안 및 R&D 키워드 간 연관관계를 분석하기 위해서는 한 문서 안에 현안 키워드와 R&D 키워드가 동시 출현하는 문서의 수집이 선행되어야 한다. 하지만 현재 현안 키워드와 R&D 키워드의 구분이 명확하지 않은 상태에서, 현안 키워드와 R&D 키워드가 혼재되어 나타나는 문서를 명시적

으로 제공하는 서비스를 찾는 것은 거의 불가능한 것으로 판단된다. 따라서 후술할 현안 및 R&D 용어 사전에 수록된 어휘를 참고하여, 다양한 서비스를 통해 제공되는 다양한 유형의 자료 중 현안과 R&D 용어간 연결고리 발굴 과정에 사용될 수 있는 자료를 식별하는 과정이 필요하다. 즉, 현안과 R&D 키워드가 동시 출현하는 문서 또는 문서의 일부분을 수집할 필요가 있으며, 이는 R&D 문서 내에서 현안 키워드가 자주 출현한 부분만을 발췌함으로써 실현 가능하다. 이러한 방식의 분석 대상 선정을 위해 국가연구기관에서 제공하는 자료를 중심으로 자료의 유형을 파악하였으며, 그 결과는 <표 1>에 제시되어 있다.

<표 1>에는 양질의 R&D 정보를 제공하는 대표적인 12개 사이트의 정보가 요약되어 있다. 하지만 이들 중 게시되어 있는 문서 수가 지나치게 적거나 발간주기가 불규칙한 일부 사이트의 경우 분석을 위한 활용도가 높지 않은 것으로 판단된다. 한편 연구개발특구도서관, 국가기술사업화종합정보망(NTB), 한국전자통신연구원(ETRI), 국가과학기술지식정보서비스(NTIS) 등의 자료들은 분석대상 문서로서 활용하기에 적합하다. 연구개

〈표 1〉 현안과 R&D 정보가 혼재되어 있는 문서 목록

| 유형 | 설명 | 활용도 |
|---|---|---|
| 유형 1 | 하나의 문서가 하나의 현안이슈를 담고 있음 | 매우 높음 |
| 유형 2 | 하나의 문서에 여러 개의 현안이슈를 담고 있으며, 장이나 절로 분리 가능 | 높음 |
| 유형 3 | 하나의 문서에 여러 개의 현안이슈를 담고 있으며, 장이나 절로 분리 불가능 | 낮음 |
| 유형 4 | 주제와 관련된 수치나 정보들을 축약해서 보여줌 | 불가능 |

| 순번 | 자료명 | 유형 | 출처 | URL | 형식 | 자료건수 | 발간주기 |
|---|---|---|---|---|---|---|---|
| 1 | 특구추천기술 | 유형1 | 연구개발특구정보도서관 | http://www.dit.or.kr | HTML | 2,809건 | 불규칙 |
| 2 | 신탁기술 | 유형1 | 연구개발특구정보도서관 | http://www.dit.or.kr | PDF | 495건 | 불규칙 |
| 3 | 사업화지정공모기술 | 유형2 | 연구개발특구정보도서관 | http://www.dit.or.kr | PDF | 91건 | 불규칙 |
| 4 | 특구최신기술 | 유형1 | 연구개발특구정보도서관 | http://www.dit.or.kr | PDF | 4,160건 | 불규칙 |
| 5 | 전체기술 | 유형1 | 국가기술사업화종합정보망 | http://www.ntb.kr | PDF | 51,106건 | 1~3일 |
| 6 | 산업기술동향 | 유형1 | 국가기술사업화종합정보망 | http://www.ntb.kr | PDF | 1,209건 | 15일 내외 |
| 7 | 신탁특허목록 | 유형1 | 연구개발특구진흥재단 | http://www.innopolis.or.kr | PDF | 495건 | 1~3일 |
| 8 | NNFC 기술소개 | 유형1 | 나노팹종합센터 | http://www.nnfc.com | HTML | 78건 | 불규칙 |
| 9 | 전자통신동향분석 | 유형1 | 한국전자통신연구원 | http://www.etri.re.kr | PDF | 1,200건 | 2개월 |
| 10 | R&D 및 정책동향 | 유형1 | 한국과총 웹진 | http://online.kofst.or.kr/ | HTML | 1,100건 | 15일 내외 |
| 11 | 보도자료 | 유형1 | 국가과학기술지식정보서비스 | http://www.ntis.go.kr/ | HTML, PDF | 121건 | 불규칙 |
| 12 | 기술동향 | 유형2 | 국가과학기술지식정보서비스 | http://www.ntis.go.kr/ | HTML, PDF | 379건 | 1~5일 |
| 13 | 언론보도 | 유형1 | K-MEG | http://www.msip.go.kr/ | HTML | 50건 | 불규칙 |
| 14 | IT R&D 정책동향 | 유형2 | 정보통신산업진흥원 | http://www.nipa.kr/ | PDF | 24건 | 1개월 |
| 15 | 정책 및 분석보고서 | 유형2 | 정보통신산업진흥원 | http://www.nipa.kr/ | PDF | 434건 | 불규칙 |
| 16 | 보고서 | 유형1 | 한국산업기술진흥원 | http://www.kiat.or.kr/ | HTML | 13,597건 | 1~2일 |
| 17 | 미래선도산업 | 유형1 | 산업통상자원부 | http://www.motie.go.kr/ | HTML, PDF | 198건 | 불규칙 |
| 18 | 분야별 동향 | 유형1 | 판교테크노밸리 | http://www.pangyotechnovalley.org/ | HTML | 691건 | 불규칙 |

발특구도서관의 경우, 연구개발사업을 진행하는 연구개발특구의 여러 기관들로부터 연결되어 있으며, 국가연구개발사업을 진행하는 연구개발특구에 관한 정보를 총체적으로 관리할 뿐 아니라, 기술시장동향, 특구기술정보, 특구기업정보 등에 대한 방대한 양의 국가 R&D 사업의 전반적인 자료를 제공하고 있다. 그리고 NTB의 경우, 추천기술, 신탁・기부・나눔 기술뿐 아니라, 산업기술동향 자료를 1~3일 주기로 제공함으로써 보다 안정적인 자료 수집이 가능하다. ETRI 역시 2개월 주기로 전자통신동향분석 보고서를 게시함으로써 사용자들에게 전자통신 분야 R&D 정보를 제공하고 있다. 또한 NTIS는 국가 R&D 사업 관련 보도자료뿐 아니라 논문, 연구보고서, 과제, 기술동향 자료들을 게시함으로써 다양한 분야의 R&D 정보를 제공하고 있다. 이처럼 위 4개 기관에서 제공하는 자료들을 그대로 활용하거나, 연구보고서 또는 논문에서 요약문, 기대효과, 결론 등의 부분만을 발췌함으로써 분석을 위한 자료를 생성할 수 있다.

### 3.2.2 현안 사전 및 R&D 용어집을 이용한 분석 대상 용어 추출

위에서 수집된 현안과 R&D 키워드가 동시 출현하는 문서들을 대상으로 현안 사전과 R&D 용어집을 이용하여 분석대상이 되는 현안 및 R&D 키워드를 도출한다. 기존의 텍스트 마이닝을 통한 연관관계 분석은 한 문서 내에 존재하는 동일한 범주의 키워드간의 연관성을 찾고자 하는 것을 전제로 한다. 그러나 본 연구에서는 서로 이질적인 현안과 R&D 키워드 간의 연관관계 분석을 수행하고자 하므로, 현안과 R&D 키워드가 동시에 출현하는 문서의 발굴을 통해 한 문서 내에서 분석대상이 되는 현안과 R&D 키워드만을 추출해야 한다. 이는 곧 분석의 품질을 향상시키고 분석 시간을 단축시키기 위해 한 문서 내에 현안과 R&D 키워드를 제외한 다른 모든 용어들을 제외시키는 정제작업이 필요함을 의미한다. 따라서 본 연구에서는 현안 사전과 R&D 용어집을 분석대상 추출을 위한 자연어처리의 Start list로 활용하고자 한다.

현안 사전은 정책보고서, 뉴스, 소셜 데이터 등의 국가현안 관련 문서들을 수집하여 토픽분석을 통해 추출된 키워드의 집합을 목록화하여 구축한다. 또한 R&D 용어집은 전문학술지 또는 과학기술 특허 등의 자료에 명시된 R&D 키워드를 식별하여 R&D 키워드 집합을 추출한 후 목록화하여 구축한다. 이 때, 텍스트 분석의 품질 향상을 위해 필요한 표준화 작업, 동음이의어, 이음동의어, 그리고 유사어를 정제하기 위한 일련의 작업은 본 연구의 범위에서 제외하였다.

이렇게 구축된 현안 사전과 R&D 용어집을 이용하여 분석대상이 되는 현안 및 R&D 키워드만을 남기고 문서 내의 다른 용어들을 제거함으로써 다음에 수행하게 될 연관관계 분석에 용이한 키워드집합을 도출한다.

### 3.2.3 현안 및 R&D 키워드 간 연관규칙 도출

앞서 설명한 과정을 통해 수집된 문서로부터 현안 키워드와 R&D 키워드 간 대응관계 도출을 위해, 우선 수집된 문서들은 현안 사전과 R&D 용어집을 활용한 정제 과정을 거치게 된다. 현안 사전 구축, R&D 용어집 구축, 그리고 정제 과정은 제 4장의 실제 사례를 통해 소개하기로 한다. 정제된 문서의 분석대상 용어 사이의 연관관계는 지지도와 신뢰도로 표현되며, 지지도에 기반한 연관규칙은 '전체 문서 중에서 키워드 A와 키워드 B가 발생할 확률은 xx%이다'와 같은 형태, 신뢰도에 기반한 연관규칙은 '키워드 A가 발생했을 때 키워드 B가 발생할 확률이 xx%이다'와 같은 형태로 기술된다. 연관규칙 도출 및 이후 프로세스의 이해를 돕기 위해, 간단한 가상 문서에 대한 분석 예를 사용하고자 한다. 현안 사전에는 (저작

권, 소프트웨어)의 2개의 용어가 포함되어 있으며, R&D 사전에는 (멀티미디어, QoS, QR-Code, 워터마크, 위조, 인증)의 6개의 용어가 포함되어 있는 것으로 가정하자. 물론 실제로는 현안 키워드와 R&D 키워드로 동시에 사용되는 용어가 다수 존재할 것으로 예상되지만, 본 예에서는 간략한 설명을 위해 이러한 경우는 나타내지 않았다. 정제 과정을 거친 가상 문서 10개(d1~d10) 각각이 포함하고 있는 현안 키워드 및 R&D 키워드가 <표 2>에 소개되어 있다.

〈표 2〉 10개의 가상 문서와 문서별 현안 및 R&D 키워드 목록

| 문서 | 현안 및 R&D 키워드 |
|---|---|
| d1 | 저작권, 워터마크 |
| d2 | 저작권, 워터마크, QR-Code, 멀티미디어 |
| d3 | 저작권, 위조, 멀티미디어 |
| d4 | 인증, 워터마크 |
| d5 | 인증, 워터마크, 위조 |
| d6 | 인증, QR-Code, 위조 |
| d7 | 멀티미디어, QoS |
| d8 | 멀티미디어, QoS, 인터넷 |
| d9 | 소프트웨어, 저작권, 위조 |
| d10 | 소프트웨어, 저작권, QR-Code, 인증 |

본 예에서는 도출된 연관규칙의 지지도에 기반한 무방향 소셜 네트워크를 구축하고자 한다. <표 2>에 제시된 문서에 대한 연관관계 분석을 통해 도출된 연관규칙 중, 지지도가 20% 이상인 규칙들의 목록은 <표 3>과 같다. 단 지지도의 경우 (항목 A → 항목 B)의 지지도와 (항목 B → 항목 A)의 지지도는 항상 동일하므로, 이를 중복 기재하지 않고 (항목 A ↔ 항목 B)로 축약하여 표현하였다.

이렇게 도출된 연관규칙의 각 항목은 현안 키워드 또는 R&D 키워드로 구성된다. 다만 현재 시점에서는 (현안 ↔ 현안, 현안 ↔ R&D, R&D ↔ R&D)의 세 가지 유형의 연관관계가 모두 존재하며, 모든 유형의 연관관계가 방법론의 다음 단계인 소셜 네트워크분석의 입력 값으로 사용된다. 규칙에 포함된 현안 및 R&D 키워드는 노드로, 연관규칙의 지지도 또는 신뢰도 값은 노드 간의 간선 값으로 사용된다. 소셜 네트워크 구축 및 이에 대한 분석 과정은 다음 절에서 설명한다.

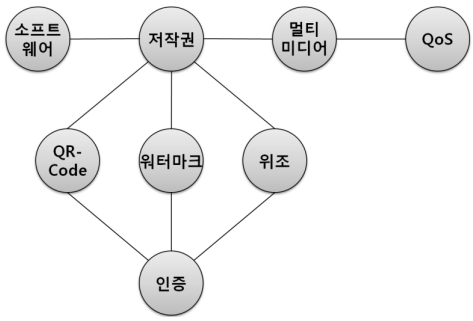
〈표 3〉 현안 및 R&D 키워드 목록에 대한 주요 연관규칙

| 연관규칙 | 유형 | 지지도 |
|---|---|---|
| 소프트웨어 ↔ 저작권 | 현안 ↔ 현안 | 20% |
| QR-Code ↔ 저작권 | 현안 ↔ R&D | 20% |
| 워터마크 ↔ 저작권 | 현안 ↔ R&D | 20% |
| 위조 ↔ 저작권 | 현안 ↔ R&D | 20% |
| 멀티미디어 ↔ 저작권 | 현안 ↔ R&D | 20% |
| 워터마크 ↔ 인증 | R&D ↔ R&D | 20% |
| 위조 ↔ 인증 | R&D ↔ R&D | 20% |
| QR-Code ↔ 인증 | R&D ↔ R&D | 20% |
| 멀티미디어 ↔ QoS | R&D ↔ R&D | 20% |

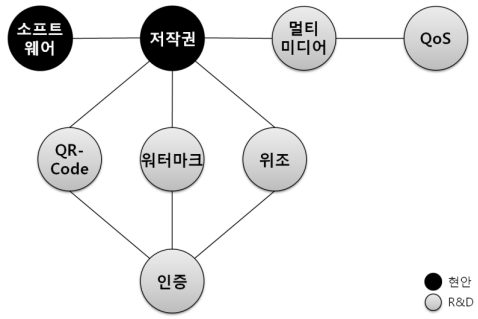
## 3.3 연관 키워드 네트워크 분석

### 3.3.1 연관 키워드 네트워크 구축

제 3.2절의 과정을 통해 도출된 현안 및 R&D 키워드 간 연관규칙 집합을 통해 각 현안에 대응되는 R&D 키워드를 식별할 수 있으나, 이는 현안과 R&D 키워드 간의 직접적인 연관성을 갖는 대응 관계만을 식별한 것으로, 다른 매개 키워드를 통한 간접적이지만 강한 연관성을 갖는 대응 관계를 파악하지 못한다는 한계를 갖는다. 이러한 한계는 <그림 4>를 통해 살펴볼 수 있다. 위의 <표 3>에서 (워터마크 ↔ 저작권), (워터마크 ↔ 인증)의 두 가지 규칙은 의미 있는 관계로 이미 파악되었지만, (저작권, 인증)의 두 키워드에 대한 연관성에 대해서는 점검하지 못하고 있다. 따라서 이러한 한계를 극복하고 보다 확장된 대응관계를 파악하기 위해, 위의 <표 3>에서 파악된 연관규칙을 소셜 네트워크로 구축하여 보다 다층적인 분석을 실시하고자 한다. 본 연구에서는 특정 현안과




〈그림 4〉 연관 키워드 네트워크 예




〈그림 5〉 속성이 부여된 연관 키워드 네트워크

관련된 R&D 키워드 간의 연결성을 분석하기 위해 지지도를 기반으로 한 무방향 키워드 네트워크를 구축하였으며, 그 결과는 <그림 4>와 같다.

하지만 <그림 4>의 소셜 네트워크의 경우, 네트워크 상에서 각 노드가 현안 키워드를 의미하는지, 아니면 R&D 키워드를 의미하는지 알 수 없다. 따라서 본 연구의 궁극적 목표인 현안 대응 R&D 키워드를 식별하기 위해서는 각 노드의 속성을 현안, R&D, 또는 현안 겸 R&D로 부여하는 작업이 필요하다. 본 예에서는 현안 겸 R&D 키워드는 사용하지 않았으므로, 현안, R&D의 속성만을 구분하여 수정한 소셜 네트워크를 <그림 5>에 제시하였다.

### 3.3.2 최단거리 및 경로 수 기반 현안 대응 R&D 매핑

위의 과정을 통해 연관 키워드간의 네트워크가 구축되면 네트워크 내의 노드간 최단거리와 최대경로 수를 기반으로 한 각각의 현안-R&D 키워드 대응 테이블을 도출한다. 네트워크 내의 두 노드 간에는 무수히 많은 연결경로가 존재할 수 있지만, 어떠한 관점에서 접근하느냐에 따라 노드간의 연결 강도를 상이하게 해석할 수 있다. 최단거리에 기반한 분석인 경우, 두 노드 사이의 최단 거리가 짧을수록 해당 노드들이 서로 높은 연결관계를 갖는 것으로 해석되며, 최대경로 수 기반의 분석인 경우 두 노드를 연결해주는 경로 수가 많을수록 더욱 신뢰성 있는 연결관계를 갖는 것으로 해석된다. 본 연구에서는 이러한 두 가지 관점을 모두 반영하여, 각 현안과 밀접한 연관성을 갖는 R&D 키워드 매핑 테이블을 도출하였다. 다음의 <표 4>와 <표 5>는 각각 <그림 4>의 네트워크에 대한 최단거리 기반 대응 테이블과 최대경로 수

〈표 4〉 최단거리 기반 현안-R&D 키워드 대응 테이블

| 키워드 | 저작권 | 소프트웨어 | QR-Code | 워터마크 | 위조 | 인증 | 멀티미디어 | QoS |
|---|---|---|---|---|---|---|---|---|
| 저작권 | 0 | 1 | 1 | 1 | 1 | 2 | 1 | 2 |
| 소프트웨어 | 1 | 0 | 2 | 2 | 2 | 3 | 2 | 3 |
| QR-Code | 1 | 2 | 0 | 2 | 2 | 1 | 2 | 3 |
| 워터마크 | 1 | 2 | 2 | 0 | 2 | 1 | 2 | 3 |
| 위조 | 1 | 2 | 2 | 2 | 0 | 1 | 2 | 3 |
| 인증 | 2 | 3 | 1 | 1 | 1 | 0 | 3 | 4 |
| 멀티미디어 | 1 | 2 | 2 | 2 | 2 | 3 | 0 | 1 |
| QoS | 2 | 3 | 3 | 3 | 3 | 4 | 1 | 0 |

〈표 5〉 최대경로 수 기반 현안-R&D 키워드 대응 테이블

| 키워드 | 저작권 | 소프트웨어 | QR-Code | 워터마크 | 위조 | 인증 | 멀티미디어 | QoS |
|---|---|---|---|---|---|---|---|---|
| 저작권 | 0 | 4 | 3 | 3 | 3 | 3 | 1 | 1 |
| 소프트웨어 | 1 | 0 | 3 | 3 | 3 | 3 | 4 | 4 |
| QR-Code | 3 | 3 | 0 | 4 | 4 | 3 | 3 | 3 |
| 워터마크 | 3 | 3 | 4 | 0 | 4 | 3 | 3 | 3 |
| 위조 | 3 | 3 | 4 | 4 | 0 | 3 | 3 | 3 |
| 인증 | 3 | 3 | 3 | 3 | 3 | 0 | 3 | 3 |
| 멀티미디어 | 1 | 4 | 3 | 3 | 3 | 3 | 0 | 1 |
| QoS | 1 | 4 | 3 | 3 | 3 | 3 | 1 | 0 |

기반 대응 테이블을 나타낸다.

<표 4>와 <표 5>에서 어둡게 표현된 영역은 현안 키워드와 R&D 키워드 간의 최단경로 거리와 최대경로 수를 나타낸다. 그 외의 부분은 현안-현안 관계, 또는 R&D-R&D 키워드 관계를 나타내므로 이후의 논의에서 제외된다. 예를 들어 <표 4>에서 현안 키워드인 “소프트웨어”의 경우, 가장 밀접한 관계가 있는 R&D 키워드는 최단경로 거리가 2인 “QR-Code”, “워터마크”, “위조”, “멀티미디어”로 파악되며, <표 5>에서 “소프트웨어”와 가장 연결 강도가 강한 R&D 키워드는 최대경로 수가 4인 “멀티미디어”와 “QoS”로 파악된다. 따라서

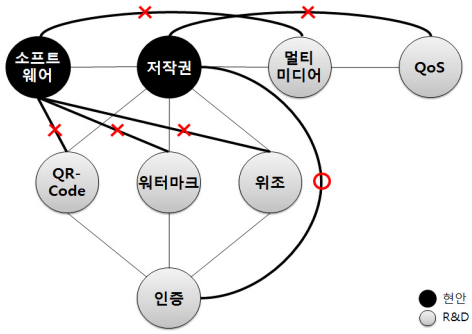
〈표 6〉 현안-R&D 키워드 간 최단경로 거리 및 최대경로 수

| 현안 | R&D | 최단경로 거리 | 최단경로 수 |
|---|---|---|---|
| 저작권 | QR-Code | 1 | 3 |
| 저작권 | 워터마크 | 1 | 3 |
| 저작권 | 위조 | 1 | 3 |
| 저작권 | 멀티미디어 | 1 | 1 |
| 저작권 | QoS | 2 | 1 |
| 저작권 | 인증 | 2 | 3 |
| 소프트웨어 | QR-Code | 2 | 1 |
| 소프트웨어 | 워터마크 | 2 | 1 |
| 소프트웨어 | 위조 | 2 | 1 |
| 소프트웨어 | 멀티미디어 | 2 | 1 |
| 소프트웨어 | QoS | 3 | 1 |
| 소프트웨어 | 인증 | 3 | 1 |

두 가지 관점을 모두 감안할 때, 현안 키워드 “소프트웨어”에 가장 밀접하게 대응되는 R&D 키워드는 “멀티미디어”임을 알 수 있다. <표 4>와 <표 5>를 동시에 요약한 결과가 <표 6>에 나타나있다.

제안 방법론은 기본적으로 최단경로의 거리가 가까운 키워드를 연관성이 높은 것으로 파악한다. 하지만 두 노드 사이에 매개 노드가 있는 경우(즉 두 노드의 최단거리가 2이상 일 때), 최대경로 수가 임계치 미만으로 나타나는 경우에는 이러한 연결 관계를 충분히 신뢰할 수 없으므로 해당 연결 관계를 인정하지 않는다. 예를 들어 <표 6>의 현안 키워드인 “저작권”을 살펴보면, (저작권 → QoS), (저작권 → 인증)의 두 가지 간접 경로를 갖는다. 이 때 최대경로 수의 임계치를 2로 적용한다면, 전자의 관계는 최대경로 수 부족으로 인해 연관성의 신뢰도가 낮은 것으로 판단된다. 한편 후자인 (저작권 → 인증)의 경우 최대경로 수가 3이므로 여러 매개 키워드(QR-Code, 워터마크, 위조)를 통해 연결되는 연결 강도가 높다고 할 수 있다. 본 예에서는 현안-R&D 키워드간의 최단경로 거리가 1인 경우, 또는 최단경로 거리가 2이면서 두 노드 사이의 최대경로 수가 2 이상인 경우만 해당 노드간 연관성이 있는 것으로 파악하고자 한다. 즉 최단경로 거리가 2일지라도 최대경로 수가 1이라면, 해당 노드간의 연관성은 신뢰할 수 없으므로 인정하지 않는다. <표 6>의 연관관계를 이

와 같은 규칙에 의해 도식화 한 결과가 <그림 6>에 나타나있다. 그림에서 점선으로 나타난 부분은 현안-현안 간 관계 또는 R&D-R&D간 관계를 나타낸다. 한편 굵게 나타난 간선은 거리가 2 이상인 간접 경로를 나타내며, 이 중 최대경로 수가 임계값을 만족시키는 경우는 O, 그렇지 않은 경우는 X로 표시하였다. 궁극적으로, <그림 6>에 나타난 간선 중 실선으로 표시된 간선(X표시 간선 제외)만이 현안 대응 R&D 매핑 테이블에 기록된다.




〈그림 6〉 현안-R&D 대응 테이블에 기반한 연관 키워드 네트워크

<그림 6>에 나타난 현안-R&D 키워드 간 대응관계를 현안 키워드 중심으로 정리한 테이블이 <표 7>에 나타나있다.

즉 이러한 모든 과정의 결과를 통해 도출한 최종 산출물은 <표 7>에 나타난 현안 대응 R&D 키워드 테이블이며, 이 테이블을 통해 사용자가 입력한 현안과 직결된 R&D 정보 패키지를 획득하는 과정은 다음 장의 사례에서 소개하도록 한다.

〈표 7〉 현안 대응 R&D 키워드 테이블

| 현안 | R&D | 현안 | R&D |
|---|---|---|---|
| 저작권 | QR-Code | 소프트웨어 | |
| | 워터마크 | | |
| | 위조 | | |
| | 멀티미디어 | | |
| | 인증 | | |

## 4. 적용 사례

### 4.1 분석 데이터 및 환경 소개

본 연구에서는 제 3장에서 제시한 현안 대응 R&D 정보 패키징 방법론을 실제로 적용하기 위해, 제 3.2.1절에서 수집한 현안-R&D 키워드 동시포함 자료 중 NTIS 사이트에서 제공하는 자료를 선정하여 실험을 수집하였다. 특히 NTIS에 게재된 논문, 과제, 연구보고서 등의 R&D 관련 자료들 중에서도 가장 체계적으로 정리가 되어있고 활용가능성이 높은 연구보고서를 실험 대상으로 채택하였다. 총 28,060건의 연구보고서 중 2010년부터 2012년까지 게재된 연구보고서의 주요 항목을 확보하였고, 대상 기간에 게재된 자료들 중 100건을 표본으로 추출하였으며, 각각의 자료에서 현안과 R&D 용어가 동시에 출현하는 요약문만을 발췌하여 이를 대상으로 방법론을 적용하였다. 키워드 추출, 연관관계 분석의 전 과정은 SAS Enterprise Miner 7.1 상에서, 연관 키워드 네트워크 구축 및 분석은 UCINET과 NetMiner 상에서 수행하였다. 또한 구축된 네트워크는 그래픽이 간결한 NodeXL을 통해 도식화하였다.

### 4.2 현안 대응 NTIS R&D 정보 패키징

#### 4.2.1 현안 및 R&D 키워드 간 연관관계 분석

NTIS에 등록된 2010년부터 2012년 사이의 연구보고서 중에서 100건을 표본으로 추출한 뒤, 요약문을 발췌하여 이를 분석에 활용하였다. 우선 분석의 품질을 향상시키고 분석 시간을 단축시키기 위한 정제작업을 수행하였으며, 이 과정에서 현안 및 R&D 키워드를 제외한 다른 용어들을 제거하기 위해 현안 사전과 R&D 용어집을 사용하였다. 현안 사전은 현안을 담고 있는 다양한 자료 중 <표 8>에 요약된 자료를 수집에 대한 토픽 분석을 통해 구축하였다.

〈표 8〉 현안 사전 구축을 위한 현안 문서 수집

| 유형 | 자료명 | 건수 |
|---|---|---|
| 정책자료 | 국회입법조사처 이슈와 논점 | 100건 |
| 뉴스 | 한겨레 신문 | 10,374건 |
| 토론 | 중앙일보 오피니언, 조선닷컴 토론마당 | 190건 |
| 컬럼 | 새로운 사회를 여는 연구원 컬럼 | 496건 |

즉 <표 8>에 정리된 자료에 대한 토픽 분석을 통해 주요 토픽을 추출하였으며, 각 토픽을 구성하고 있는 어휘들을 모두 취합하여 이를 현안 사전으로 구축하였다. 토픽 분석은 SAS Enterprise Miner 7.1의 모듈인 Text Miner에서 수행하였으며, 그 예로 정책자료에 대한 토픽 분석 결과를 <그림

| 토픽 ID | 문서 임계치 | 용어 임계치 | 토픽 | 용어 수 | 문서 수 |
|---|---|---|---|---|---|
| 1 | 0.363 | 0.045 | 북한,대북,핵실험,핵,북 | 164 | 7 |
| 2 | 0.313 | 0.041 | 환율,엔,환화,강세,인하 | 210 | 7 |
| 3 | 0.312 | 0.040 | 헌법재판소,위헌,결정,해석,변형 | 193 | 6 |
| 4 | 0.307 | 0.041 | 학교,교육과정,학생,교육,영어 | 209 | 8 |
| 5 | 0.298 | 0.039 | 약관,설정,채무자,대출,담보 | 164 | 5 |
| 6 | 0.290 | 0.039 | 심사,예산,의회,하원,결산 | 194 | 4 |
| 7 | 0.291 | 0.038 | 근로자,근로,고용,임금,새 | 257 | 6 |
| 8 | 0.283 | 0.037 | 협정,협상,철레,개국,체결 | 221 | 5 |
| 9 | 0.252 | 0.036 | 정당,선거,교육감,지방,기초 | 189 | 4 |
| 10 | 0.281 | 0.036 | 테러,지역,안보,일본,미 | 234 | 7 |
| 11 | 0.259 | 0.036 | 과학,기술,과학,혁신,연구,기초 | 217 | 5 |
| 12 | 0.262 | 0.035 | 업종,대기업,중소기업,컨고,선정 | 221 | 4 |
| 13 | 0.251 | 0.035 | 인수위,조직,개편,대통령직,당선인 | 218 | 2 |
| 14 | 0.246 | 0.035 | 건축물,건축,공사,안전관리,중단 | 215 | 8 |
| 15 | 0.247 | 0.035 | 거래,금융,조세,현금,자금 | 227 | 4 |
| 16 | 0.249 | 0.034 | 게임,물,청소년,인터넷,평가 | 235 | 5 |
| 17 | 0.240 | 0.034 | 의료,환자,진료,병원,인력 | 212 | 4 |
| 18 | 0.225 | 0.034 | 전력,에너지,수급,발전소,설비 | 221 | 4 |
| 19 | 0.240 | 0.034 | 여성,고용,근로자,관리자,임금 | 219 | 6 |

〈그림 7〉 현안 사전 구축을 위한 토픽 분석 화면 예

| | F1 | F2 |
|---|---|---|
| 1 | text01 | 연구개발의 목적 및 필요성....인터넷을 근간으로 하는 정보화 사회에서 사이버범죄… |
| 2 | text02 | 요 약 문.. Ⅰ. 제 목..DCI 규격을 준수하는 디지털시네마 배급관리 및 저작권 보호를… |
| 3 | text03 | 연구의..목적 및 내용..본 연구에서는 GIS 콘텐츠의 불법 복제 및 유통 방지와 안전성… |
| 4 | text04 | Ⅰ. 제 목..CODEX콘텐츠 제작을 위한 진화형 RPG 캐릭터 AI 기술개발.. Ⅱ. 기술개발… |
| 5 | text05 | 최종보고 요약서..연구소명 모바일 플랫폼 개발 연구소..과제구분 ■ 신규 과제 업그… |
| 6 | text06 | 요 약 서 (초 록)..(국문) SOA 기반 양방향 동영상 응용 서비스 기술 표준개발..과 제… |
| 7 | text07 | 요 약 문.. Ⅰ. 연구목표 및 내용..본 연구에서는 윌 ? 바이러스 및 해킹 등의 사이버 침… |
| 8 | text08 | 연구결과 요약문..한글요약문..양식 A202..연구의..목적 및 내용..본 연구에서는 인문… |
| 9 | text09 | 요 약 문.. Ⅰ. 제 목..디지털 문화콘텐츠 융복합 서비스를 위한 시맨틱 웹 매쉬업 플랫… |
| 10 | text10 | 〈 연구결과 요약문 〉..중심어..디지털 워터마킹 건축설계도면..CAD 저작권 보호..D… |

〈그림 8〉 분석 대상 키워드 추출 이전의 자료 원문(일부)

| 용어 | 역할 | 속성 | 상태 ▲ | 가중 | 가져온 빈도 | 빈도 | 가져온 문서 수 |
|---|---|---|---|---|---|---|---|
| 통합 | | 알파 | 유지 | 0.465 | 52 | 52 | 18 |
| 소프트웨어 | | 알파 | 유지 | 0.514 | 54 | 54 | 18 |
| 스마트 | | 알파 | 유지 | 0.466 | 73 | 73 | 18 |
| 산업 | | 알파 | 유지 | 0.432 | 30 | 30 | 17 |
| 상황 | | 알파 | 유지 | 0.453 | 41 | 41 | 17 |
| 비용 | | 알파 | 유지 | 0.444 | 27 | 27 | 17 |
| 모니터링 | | 알파 | 유지 | 0.458 | 45 | 45 | 17 |
| 발생 | | 알파 | 유지 | 0.466 | 56 | 56 | 17 |
| 단계 | | 알파 | 유지 | 0.457 | 34 | 34 | 17 |
| 변화 | | 알파 | 유지 | 0.487 | 54 | 54 | 17 |
| 공동 | | 알파 | 유지 | 0.414 | 26 | 26 | 17 |
| 효율적 | | 알파 | 유지 | 0.404 | 23 | 23 | 17 |
| 치료 | | 알파 | 유지 | 0.487 | 39 | 39 | 17 |

〈그림 9〉 현안 및 R&D 키워드만으로 구성된 파싱 결과(일부)

7>에 제시하였다. 또한 R&D 용어집은 대상 기간에 NTIS에 등록된 총 8,501건의 연구보고서를 통해 구축하였다. 즉 이들 자료에 명시된 키워드 중 15,873개의 R&D 키워드 집합을 목록화하여 구축하였다.

이렇게 구축된 현안 사전과 R&D 용어집에 등록된 어휘의 합집합을 키워드 추출 시 Start List로 적용함으로써, 현안 및 R&D 키워드만으로 구성된 파싱 결과를 얻을 수 있다. <그림 8>은 Start List를 적용하기 이전의 자료 원문 일부를 보이고 있으며, 이에 대한 파싱을 통해 <그림 9>의 결과를 얻을 수 있다. 또한 이렇게 추출된 현안 및 R&D 키워드의 집합에 대한 연관관계 분석을 통해 도출한 10,000개의 연관규칙 중 일부가 <표 9>에 나타나있다.

〈표 9〉 도출된 연관규칙(일부)

| Rule | Conf. | Sup. | Lift |
|---|---|---|---|
| it ⇒ 보호 | 54.55 | 5.94 | 2.50 |
| it ⇒ 센서 | 54.55 | 5.94 | 4.59 |
| it ⇒ 응용 | 54.55 | 5.94 | 2.5 |
| it ⇒ 제어 | 54.55 | 5.94 | 4.24 |
| sw ⇒ 구조 | 100.00 | 3.96 | 4.59 |
| sw ⇒ 데이터 | 100.00 | 3.96 | 2.81 |
| sw ⇒ 특허 | 100.00 | 3.96 | 2.97 |
| 건강 ⇒ 검증 | 71.43 | 4.95 | 5.15 |
| 건강 ⇒ 모니터링 | 57.14 | 3.96 | 3.39 |
| 건강 ⇒ 임상 | 57.14 | 3.96 | 5.25 |
| 저작권 ⇒ 데이터 | 80.00 | 7.92 | 2.24 |
| 저작권 ⇒ 디지털 | 50.00 | 4.95 | 3.88 |

### 4.2.2 연관 키워드 네트워크 분석

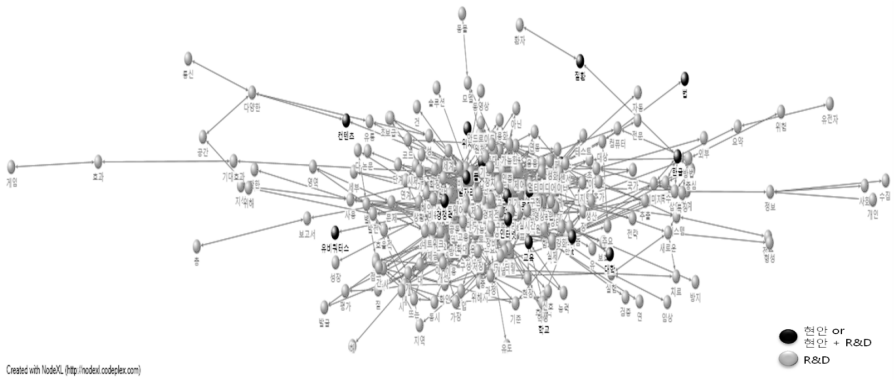
제 4.2.1절에서 도출된 현안-R&D 키워드 간의 연관규칙 소셜 네트워크로 구축하여 분석한다. 본 연구에서는 특정 현안 키워드와 관련된 R&D 키워드 간의 연관 방향성 보다는, 그와 관련된 R&D 키워드 간의 관계, 즉 특정 현안과 관련된 R&D 키워드 간의 연결 강도를 통하여 연관성이 높은 현안과 R&D 키워드 간의 연관성을 파악하기 위해 지지도 기반의 연관 키워드 네트워크를 구축하였다. 앞에서 도출한 10,000개의 규칙 중 지지도가 9% 이상인 규칙 1,000개를 추출하여 네트워크를 구축하였으며, 노드별 속성을 부여해 각 노드가 '현안', 'R&D', '현안 + R&D'의 총 3가지 중에서 어떤 것을 의미하는지를 식별 가능하도록 속성을 부여해 시각적으로 구분하였다. 연관규칙 1,000개를 통해 구축된 네트워크를 NodeXL을 통해 도식화 한 결과가 <그림 10>에 나타나있다.

<그림 10>의 네트워크는 임계치 이상의 지지도를 갖는 1,000개의 규칙만을 도식화하고 있음에도, 그 관계가 매우 복잡하여 결과 분석이 어려운 측면이 있다. 따라서 결과 해석을 위해 위 네트워크 중 100개+의 규칙만을 추출하여 간략화한 결과가 <그림 11>에 나타나있다.

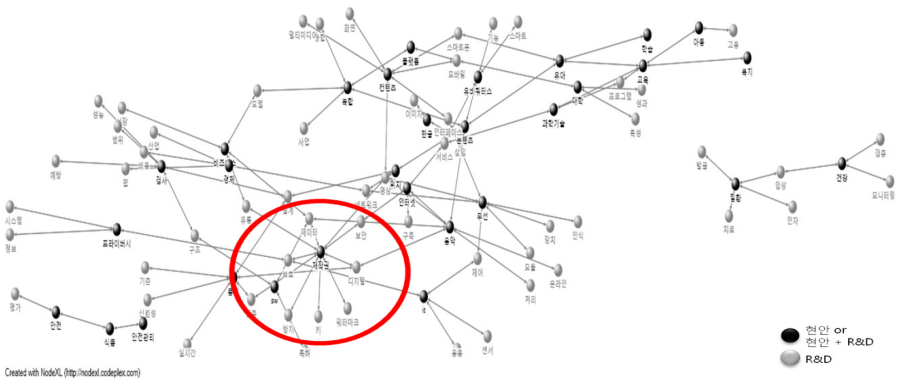
<그림 11>에서 원으로 표시된 부분을 살펴보면 '저작권'이라는 현안 키워드에 '디지털', '키', '워터마크', '보호', '보안'이라는 R&D 키워드가 연결되어 있는 것을 쉽게 확인 할 수 있다. 이처럼 연관 키워드간 네트워크가 구축되면, 노드간 최단경로 거리(<그림 12>) 및 최대경로 수 분석을 통해 현안-R&D 키워드 대응 테이블을 도출할 수 있다. <그림 11>의 네트워크의 분석 결과인 최단경로 거리와 최대경로 수를 요약한 결과가 <표 10>에 나타나있다.

최종 현안-R&D 대응 테이블 도출을 위한 마지막 단계는, <표 10>의 테이블로부터 각 현안별 주요 R&D 키워드를 선정하는 것이다. 이 과정은 임계치 이상의 최대경로 수를 갖는 두 노드 쌍을 최단경로 거리가 짧은 순으로 선택함으로써 이루어진다. 본 적용 사례에서는 최단경로 거리가 1 또는 2인 경우는 최대경로 수에 무관하게 연관성이 높은 것으로 해석하였으며, 최단경로 거리가

3 이상인 경우는 최대경로 수가 2 이상인 관계만 연관성을 인정하였다. 이 과정에서 <표 10>에 나타난 총 1,708개의 관계 중 1,020개의 관계가 제거되었다. 남은 관계 중 각 현안 별로 최단경로 거리가 가까운 상위 5개 R&D 키워드를 대응시킨 결과가 <표 11>에 제시되어 있다.




〈그림 10〉 연관규칙 1,000개로 구성된 연관 키워드 네트워크(지지도 9% 이상)




〈그림 11〉 연관규칙 100개로 구성된 연관 키워드 네트워크(지지도 9% 이상)

| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | it | sw | 건강 | 검사 | 검증 | 경제 | 고용 | 과학기술 | 교육 | 구조 |
| 1 | it | 0.000 | 4.000 | 11.000 | 6.000 | 11.000 | 4.000 | 9.000 | 6.000 | 7.000 | 5.000 |
| 2 | sw | 4.000 | 0.000 | 11.000 | 2.000 | 11.000 | 6.000 | 7.000 | 5.000 | 5.000 | 1.000 |
| 3 | 건강 | 11.000 | 11.000 | 0.000 | 11.000 | 1.000 | 11.000 | 11.000 | 11.000 | 11.000 | 11.000 |
| 4 | 검사 | 6.000 | 2.000 | 11.000 | 0.000 | 11.000 | 4.000 | 7.000 | 6.000 | 5.000 | 1.000 |
| 5 | 검증 | 11.000 | 11.000 | 1.000 | 11.000 | 0.000 | 11.000 | 11.000 | 11.000 | 11.000 | 11.000 |
| 6 | 경제 | 4.000 | 6.000 | 11.000 | 4.000 | 11.000 | 0.000 | 7.000 | 6.000 | 5.000 | 5.000 |
| 7 | 고용 | 9.000 | 7.000 | 11.000 | 7.000 | 11.000 | 7.000 | 0.000 | 3.000 | 2.000 | 8.000 |
| 8 | 과학기술 | 6.000 | 5.000 | 11.000 | 6.000 | 11.000 | 6.000 | 3.000 | 0.000 | 1.000 | 6.000 |
| 9 | 교육 | 7.000 | 5.000 | 11.000 | 5.000 | 11.000 | 5.000 | 2.000 | 1.000 | 0.000 | 6.000 |
| 10 | 구조 | 5.000 | 1.000 | 11.000 | 1.000 | 11.000 | 5.000 | 8.000 | 6.000 | 6.000 | 0.000 |
| 11 | 구축 | 5.000 | 4.000 | 11.000 | 6.000 | 11.000 | 6.000 | 6.000 | 3.000 | 4.000 | 5.000 |
| 12 | 기능 | 7.000 | 5.000 | 11.000 | 5.000 | 11.000 | 5.000 | 6.000 | 3.000 | 4.000 | 6.000 |
| 13 | 기준 | 5.000 | 5.000 | 11.000 | 3.000 | 11.000 | 5.000 | 8.000 | 6.000 | 6.000 | 4.000 |
| 14 | 네트워크 | 3.000 | 5.000 | 11.000 | 3.000 | 11.000 | 1.000 | 6.000 | 5.000 | 4.000 | 4.000 |
| 15 | 대학 | 6.000 | 6.000 | 11.000 | 7.000 | 11.000 | 6.000 | 7.000 | 6.000 | 5.000 | 7.000 |
| 16 | 데이터 | 3.000 | 1.000 | 11.000 | 3.000 | 11.000 | 5.000 | 6.000 | 4.000 | 4.000 | 2.000 |
| 17 | 디지털 | 3.000 | 3.000 | 11.000 | 3.000 | 11.000 | 5.000 | 6.000 | 4.000 | 4.000 | 4.000 |
| 18 | 멀티미디어 | 5.000 | 5.000 | 11.000 | 7.000 | 11.000 | 5.000 | 6.000 | 5.000 | 4.000 | 6.000 |
| 19 | 모니터링 | 11.000 | 11.000 | 1.000 | 11.000 | 2.000 | 11.000 | 11.000 | 11.000 | 11.000 | 11.000 |
| 20 | 모델 | 5.000 | 3.000 | 11.000 | 5.000 | 11.000 | 5.000 | 6.000 | 5.000 | 4.000 | 4.000 |
| 21 | 모듈 | 3.000 | 5.000 | 11.000 | 5.000 | 11.000 | 3.000 | 8.000 | 6.000 | 6.000 | 6.000 |
| 22 | 모바일 | 5.000 | 5.000 | 11.000 | 6.000 | 11.000 | 5.000 | 8.000 | 5.000 | 4.000 | 6.000 |
| 23 | 무선 | 2.000 | 4.000 | 11.000 | 4.000 | 11.000 | 2.000 | 7.000 | 5.000 | 5.000 | 5.000 |
| 24 | 발굴 | 11.000 | 11.000 | 3.000 | 11.000 | 4.000 | 11.000 | 11.000 | 11.000 | 11.000 | 11.000 |
| 25 | 방지 | 3.000 | 3.000 | 11.000 | 5.000 | 11.000 | 5.000 | 8.000 | 5.000 | 6.000 | 4.000 |
| 26 | 범위 | 7.000 | 3.000 | 11.000 | 1.000 | 11.000 | 5.000 | 8.000 | 7.000 | 6.000 | 2.000 |
| 27 | 보안 | 3.000 | 3.000 | 11.000 | 5.000 | 11.000 | 5.000 | 6.000 | 3.000 | 4.000 | 4.000 |

〈그림 12〉 100개 규칙에 대한 연관 네트워크의 최단경로 거리 테이블

〈표 10〉 현안 키워드와 R&D 키워드 간 최단경로 거리 및 최대경로 수(일부)

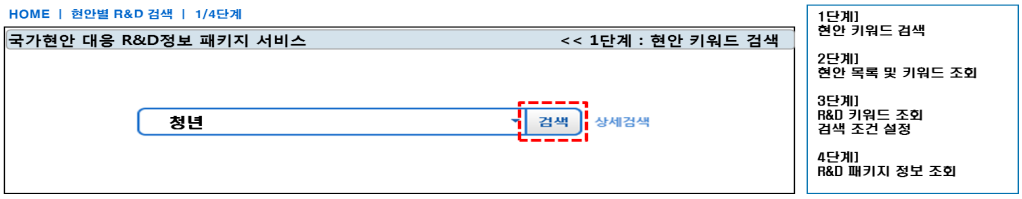
| 현안 | R&D | 최단경로 거리 | 최단경로 수 | 현안 | R&D | 최단경로 거리 | 최단경로 수 |
|---|---|---|---|---|---|---|---|
| it | 보호 | 1 | 1 | 교육 | 스마트폰 | 2 | 1 |
| it | 센서 | 1 | 1 | 교육 | 콘텐츠 | 2 | 1 |
| it | 제어 | 1 | 1 | 교육 | 학습 | 2 | 1 |
| sw | 구조 | 1 | 1 | 저작권 | 보안 | 1 | 1 |
| sw | 데이터 | 1 | 1 | 저작권 | 보호 | 1 | 1 |
| sw | 특허 | 1 | 1 | 저작권 | 디지털 | 1 | 1 |
| 건강 | 모니터링 | 1 | 1 | 저작권 | 워터마크 | 1 | 1 |
| 건강 | 임상 | 1 | 1 | 저작권 | 키 | 1 | 1 |
| 건강 | 질환 | 2 | 1 | 저작권 | 무선 | 1 | 1 |
| 건강 | 발굴 | 3 | 1 | 저작권 | 프라이버시 | 2 | 1 |
| 건강 | 인자 | 3 | 1 | 저작권 | it | 2 | 1 |
| 건강 | 치료 | 3 | 1 | 저작권 | 인터넷 | 2 | 1 |
| 교육 | 아동 | 1 | 1 | 저작권 | 서비스 | 3 | 1 |
| 교육 | 서비스 | 1 | 1 | 저작권 | 센서 | 3 | 1 |
| 교육 | 고용 | 1 | 1 | 저작권 | 제어 | 3 | 1 |

〈표 11〉 최종 현안-R&D 대응 테이블

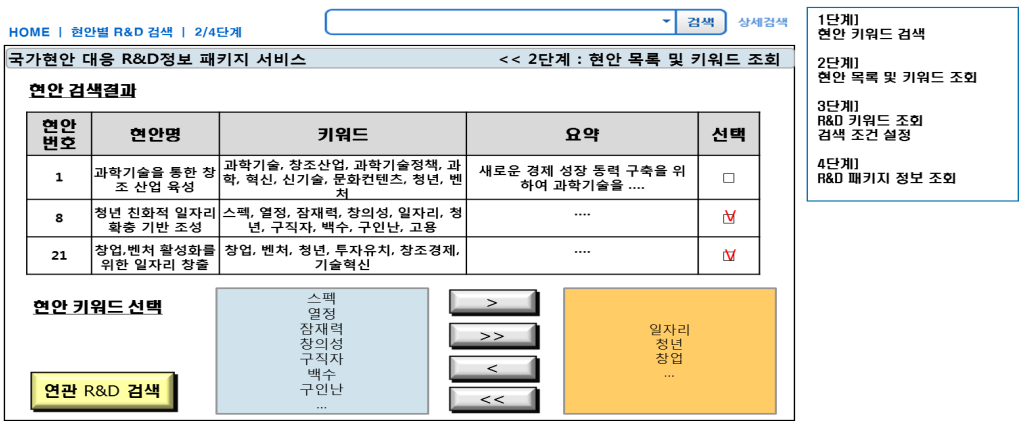
| 현안 | R&D | 현안 | R&D | 현안 | R&D | 현안 | R&D | 현안 | R&D | 현안 | R&D |
|---|---|---|---|---|---|---|---|---|---|---|---|
| IT | 보호 | 경제 | 네트워크 | 무선 | 영상 | 아동 | 고용 | 음악 | 콘텐츠 | 컨텐츠 | 영상 |
| | 제어 | | 비용 | | 네트워크 | | 교육 | | 데이터 | | 모바일 |
| | 센서 | | 산업 | | 인터페이스 | | 과학기술 | | 영상 | | 스마트폰 |
| | 응용 | | 웹 | | 제어 | | 프로그램 | | 디지털 | | 인터페이스 |
| | 무선 | | 무선 | | 인식 | | | | 인터넷 | | 멀티미디어 |
| sw | 구조 | 과학기술 | 교육 | 복지 | 교육 | 위치 | 설계 | 인터넷 | 서비스 | 품질 | 디지털 |
| | 데이터 | | 서비스 | | 과학기술 | | 콘텐츠 | | 음악 | | 설계 |
| | 특허 | | 프로그램 | | 아동 | | 네트워크 | | 보안 | | 인증 |
| | 검사 | | 유비쿼터스 | | 프로그램 | | 무선 | | 구축 | | 기준 |
| | 저작권 | | 인터넷 | | | | 유비쿼터스 | | 콘텐츠 | | 신뢰성 |
| 건강 | 검증 | 교육 | 과학기술 | 비지니스 | 데이터 | 유비쿼터스 | 서비스 | 저작권 | 데이터 | 프라이버스 | 보호 |
| | 모니터링 | | 프로그램 | | 모델 | | 콘텐츠 | | 영상 | | 시스템 |
| | 임상 | | 아동 | | 유통 | | 기능 | | 디지털 | | 정보 |
| | 질환 | | 스마트폰 | | 시장 | | 스마트 | | 보안 | | IT |
| | | | 콘텐츠 | | 저작권 | | 위치 | | 인증 | | 저작권 |
| 검사 | 구조 | 대학 | 모바일 | 식품 | 안전 | 유아 | 콘텐츠 | 질환 | 발굴 | 학습 | 교육 |
| | 설계 | | 성과 | | 안전관리 | | 교육 | | 인자 | | 스마트폰 |
| | 범위 | | 특성 | | 평가 | | 스마트폰 | | 임상 | | 콘텐츠 |
| | 성능 | | 컨텐츠 | | | | 학습 | | 치료 | | |
| | 예방 | | 플랫폼 | | | | 컨텐츠 | | 건강 | | |

### 4.2.3 현안 대응 R&D 정보 패키징 프로토타입

본 부절에서는 위의 모든 과정을 통한 현안 대응 R&D 정보 패키징 결과를 사용자들에게 어떠한 방식으로 서비스할 지에 대한 프로토타입을 제시한다. 이해를 돕기 위해 사용자가 검색 창에 특정 현안 키워드를 입력하는 가상 경우를 예로 들어 총 4단계의 과정으로 설명하기로 한다(<그림 13>~<그림 16>). 사용자가 '청년실업'이라는 현안에 대한 관련 R&D 정보를 찾고자 할 때, 1단계로 검색 창에 '청년'을 입력하면, 2단계에서 청년과 관련된 현안 목록이 제시되고, 관련 키워드와 요약이 함께 나타난다. 이 때, 사용자가 본인이 원




〈그림 13〉 1단계-현안 키워드 검색




〈그림 14〉 2단계-현안 목록 및 키워드 조회


〈그림 15〉 3단계 - R&D 키워드 조회 및 검색 조건 설정


〈그림 16〉 4단계 - R&D 패키지 정보 조회

하는 현안 목록을 선택하고 연관 R&D 검색 버튼을 클릭하면, 3단계에서 선택한 현안과 관련된 R&D 키워드 Map이 제시된다. 키워드 Map은 각 R&D 키워드별로 서로 다른 색상으로 구분되어 표시되며, 키워드의 크기가 클수록 연관성이 높은, 즉 최단경로 거리가 가까운 키워드를 의미한다. 이에 대해 구체적 검색 조건 제시함으로써 마지막 4단계에서 최종적으로 주어진 현안에 대응되는 R&D 정보 패키지를 획득할 수 있다.

## 5. 결 론

빅데이터 분석을 통해 국가현안이나 이슈를 발굴하고, 이와 관련된 R&D 정보를 획득하고자 하는 시도가 꾸준히 있어왔다. 이러한 시도는, 많은 사용자들이 특정 시기에 이슈가 되는 현안에 대해 기술 문서, R&D 문서 등을 통해 보다 전문적이고 체계적인 정보를 얻기 위한 수요가 점차 증가하고 있음을 감안할 때 매우 중요할 뿐 아니라 활용 가능성도 매우 높은 시도로 보인다. 하지만 대부분의 사용자가 현안 키워드에는 익숙하지만 관련 R&D 키워드에는 익숙하지 않기 때문에, 현재 제공되는 서비스를 통해 원하는 자료를 획득하기가 매우 어려운 현실이다. 이러한 한계의 근본 원인은 사용자가 인식하고 있는 현안 키워드와, 실제 이에 대응되는 R&D 분야에서 사용되는 키워드 간에 이질성이 존재하기 때문이다. 이러한 한계를 극복하기 위해 본 연구에서는 국가 현안 용어와 R&D 용어를 연결할 수 있는 중간 장치를 고안하고, 이로부터 국가 현안에 대응되는 R&D 문서를 효율적으로 패키징하기 위한 방법론을 제시하였다. 제안 방법론은 하나의 공통된 도메인에서 단어를 추출하고 이에 대한 연관성을 분석한 기존 연구와 달리, 서로 다른 용어를 사용하는 현안 키워드 풀과 R&D 키워드 풀의 이질성을 극복하기 위한 중간 장치를 구축하기 위한 방안을 제시하였다는 측면에서 그 기여가 인정될 수 있다.

또한 본 연구에서는 제안 방법론의 실제 적용 가능성을 평가하기 위해 다양한 매체의 현안 자료, 그리고 NTIS에 등록된 방대한 양의 R&D 자료에 대한 분석을 수행하였다. 즉 정책자료, 뉴스, 토론, 컬럼 등의 현안 자료에 대한 토픽 분석을 통해 현안 사전을 구축하고, NTIS에 등록된 연구보고서의 키워드 목록을 이용하여 R&D 사전을 구축하였다. 현안 사전과 R&D 사전을 활용하여 연구보고서에 대한 토픽 분석, 연관성 분석, 소셜 네트워크 분석을 실시하였으며, 그 결과로 현안 대응 R&D 키워드 테이블을 도출하였다. 도출된

테이블을 활용하기 위한 서비스 방안으로 4단계로 구분된 프로토타입을 제시하였으며, 이러한 전체 과정을 통해 특정 현안과 관련하여 익숙하지 않은 R&D 문서를 검색하고자 하는 사용자에게, 다양한 R&D 자료를 효과적으로 패키징하여 제공해 줄 수 있을 것으로 기대한다.

학술적 측면에서 본 연구는 서로 이질적인 용어 풀간의 대응관계를 파악했다는 점에서 그 기여가 인정될 수 있다. 즉 현안 용어간의 연관 현안 분석 그리고 R&D 용어간의 연관 R&D 분석에 대해서는 많은 연구가 이미 이루어졌지만, 특정 현안에 대응되는 R&D 용어를 파악하기 위한 연구는 상대적으로 찾아보기 어렵다. 본 연구에서 이루어진 시도는 현안과 R&D 용어 관계뿐 아니라, 서로 이질적인 용어 풀간의 대응관계 파악을 위한 유사 연구에서도 활용도가 높을 것으로 기대된다. 실무적 측면에서 본 연구는 특정 현안과 직/간접적으로 관련이 있는 R&D 자료를 제공하는 서비스의 품질을 획기적으로 개선시킬 수 있을 것으로 기대한다. 즉 제안 방법론을 R&D 자료 서비스 기관의 시스템에 적용함으로써, 특정 분야의 전문가가 아닌 일반 사용자도 관심 현안에 대한 R&D 정보를 쉽게 획득할 수 있을 것으로 기대한다.

본 연구의 후속연구에서 반드시 다루어져야 할 사항은 다음과 같다. 본 연구는 특정 현안에 대응되는 R&D 정보를 패키징하기 위한 방법론을 제안하는 것을 목적으로 수행되었기 때문에, 방법론의 이해를 돕기 위해 수행한 실험 과정에서 이루어진 일부 의사결정이 명확한 근거에 기반하고 있지 않은 측면이 있다. 예를 들면 연관규칙 추출을 위한 지지도 선택, 추출 표본의 개수, 분석 대상 문서의 선정 등이 그 예이다. 향후 연구에서는 엄밀한 기준에 의해 선정된 데이터에 대해 더욱 정교한 실험을 수행해야 하며, 이 과정에서 실험에 필요한 각 단계마다의 의사결정에 대한 근거가 도출되어야 한다. 또한 실험 결과의 공신력을 높이기 위해서는 텍스트 분석의 품질 향상을 위해 필요한 표준화 작업, 동음이의어, 이음동의어, 그리고 유사어를 정제하기 위한 일련의 작업이 객관적인 도구에 의해 수행되어야 한다. 또한 본 연구의 실험에서는 실험 도구의 한계로 인해 표본 자료에 대한 제한된 분석을 실시하였으나, 실제 서비스 적용을 위해서는 보다 다양한 사이트에 등록된 방대한 양의 전체 자료에 대한 추가 실험이 필요한 것으로 판단된다.

## 참 고 문 헌

[1] 강은영, 곽기영, “Managing Duplicate Memberships of Websites : An Approach of Social Network Analysis”, *지능정보연구*, 제17권 제1호, 2011, pp. 153-169.

[2] 곽기영, 소셜 네트워크분석, 서울 : 청람, 2013 (Forthcoming).

[3] 권이남, 김재수, “국가 R&D정보 참조연계 서비스 시스템 구축에 관한 연구”, *한국콘텐츠학회논문지*, 제8권 제1호, 2008.

[4] 김경재, 안현철, “개선된 데이터마이닝 기술에 의한 웹 기반 지능형 추천시스템 구축”, *Journal of Information Technology Applications and Management*, 제12권 제3호, 2005, pp. 41-56.

[5] 김남규, “장바구니 크기가 연관규칙 척도의 정확성에 미치는 영향”, *경영정보학연구*, 제18권 제2호, 2008, pp. 95-114.

[6] 김문수, 이학연, 최창우, 이성룡, 최경일, 전진우, “국가연구개발 성과추적평가관리 시스템 모형 및 활용”, *기술혁신학회지*, 제11권 제4호, 2008, pp. 613-638.

[7] 김용학, 사회 연결망 분석, 박영사, 2003.

[8] 김인현, “빅데이터 가치와 도입 전략”, 2012 Big Data 검색 분석 기술 Insight, 2012.

[9] 류범종, "국가 R&D 성과정보의 효율적인 관리 및 유통체제 구축에 관한 연구", *한국문헌정보학회지*, 제37권 제4호, 2003, pp. 223-240.
[10] 류범종, 최기석, "국가 R&D 지식정보관리시스템 구축에 관한 연구 - 연구기획 및 관리를 중심으로", *한국문헌정보학회지*, 제38권 제1호, 2004, pp. 281-301.
[11] 류법모, 김현진, 김현기, 박상규, "심층 언어 분석 기반 소셜미디어 이슈 탐지 및 모니터링 기술", *정보과학회지*, 제30권 제6호, 2012, pp. 47-58.
[12] 박우창, 승현우, 용환승, 데이터마이닝 : 개념 및 기법, 서울 : 자유아카데미, 2003.
[13] 손동원, 사회 네트워크 분석, 경문사, 2002.
[14] 신성호, 윤영준, 양명석, 김진만, 손강렬, "데이터 품질을 고려한 국가 R&D 정보 데이터베이스의 통합 사례 연구 - NTIS 데이터베이스 통합 사례", *한국컴퓨터정보학회 논문집*, 제16권 제6호, 2011. 6.
[15] 안현철, 한인구, 김경재, "연관규칙기법과 분류모형을 결합한 상품추천시스템 : G인터넷쇼핑몰의 사례", *Information System Review*, 제8권 제1호, 2006, pp. 181-201.
[16] 양명석, 윤영준, 신성호, 김진만, 손강렬, "국가 R&D 참여인력 DB를 활용한 전문가 검색 대행서비스 구축", 한국인터넷정보학회 추계학술발표대회 논문집, 제11권 제2호, 2010.
[17] 윤성준, "데이터마이닝 기법을 통한 백화점의 고객이탈예측 모형 연구", *한국마케팅저널*, 제6권 제4호, 2005, pp. 45-72.
[18] 이부형, "빅데이터의 생성과 새로운 사업 기회 창출", 현대경제연구원, 2012.
[19] 이연정, 김경재, "다중모형조합기법을 이용한 상품추천시스템", *지능정보연구*, 2013.
[20] 조인동, 김남규, "소셜 네트워크와 데이터마이닝 기법을 활용한 학문 분야 중심 및 융합 키워드 추천 서비스", *지능정보연구*, 제17권 제1호, 2011, pp. 127-138.
[21] 최광선, "SNS 시대의 하이브리드 빅데이터 분석 기술 및 사례", 2012 Big Data 검색 분석기술 Insight, 2012.
[22] 최창현, "조직의 비공식 연결망에 관한 연구 : 사회연결망 분석의 적용", *한국사회와 행정연구*, 제17권 제1호, 2006, pp. 1-23.
[23] 허준, 김종우, "혼합 데이터 마이닝 기법인 불일치 패턴 모델의 특성 연구", *Journal of Information Technology Applications and Management*, 제15권 제1호, 2008, pp. 225-242.
[24] 황인수, "정보검색에서 웹마이닝을 이용한 동적인 질의확장에 관한 연구", *Journal of Information Technology Applications and Management*, 제11권 제2호, 2004, pp. 227-237.
[25] Agrawal, R. and Srikant, R., "Fast Algorithms for Mining Association Rules", International Conference on Very Large Data Bases, Santiago, Chile, 1994, pp. 487-499.
[26] Albright, R., "Taming Text with the SVD", SAS Institute Inc., 2006.
[27] Fan, W., Wallace, W., Rich, S., and Zhang, Z., "Tapping the Power of Text Mining", *Communications of the ACM*, Vol. 49, No. 9, 2006, pp. 76-82.
[28] Freeman, L. C., Social Network Analysis, SAGE, 2008.
[29] Gartner Inc., "2012 Hype Cycle for Emerging Technologies", Gartner Inc., 2012.
[30] Han, J. and Kamber, M., Data Mining: Concepts and Techniques, (3rd ed.), Morgan Kaufmann Publishers, 2011.
[31] Hearst, M. A., "Untangling Text Data Mining", in Proceedings of the 37th ACL, 1999.
[32] IDC, "2011 IDC Digital Universe Study :

Big Data Is Here, Now What?", IDC, 2011

[33] IDC, "IDC Releases First Worldwide Big Data Technology and Services Market Forecast, Shows Big Data as the Next Essential Capability and a Foundation for the Intelligent Economy", IDC, March, 2012.

[34] Kauffman, S., The Origins of Order, Oxford University Press, 1993.

[35] McKinsey Global Institute, "Big Data : The next Frontier for Innovation, Competition, and Productivity", McKinsey and Company, 2011.

[36] Metzler, D., Bernstein, Y., Crofit, W. B., Moffat, A., and Zobel, J., "Similarity Measures for Tracking Information Flow", in Proceedings of CIKM, Bremen, Germany, 2005.

[37] Mooney, R. J. and Bunescu, R., "Mining Knowledge from Text using Information Extraction", *ACM SIGKDD Explorations*, Vol. 7, No. 1, 2006, pp. 3-10.

[38] O'Reilly Radar Team, Big Data Now : Current Perspectives from O'Reilly Radar, O'Reilly, 2011.

[39] Rijsbergen, C. J. V., Information Retrieval, (2nd ed.), Butterworth, London, 1979.

[40] Salton, G., Wong, A., and Yang, C. S., "A Vector Space Model for Automatic Indexing", *Communications of the ACM*, Vol. 18, No. 11, 1975, pp. 613-620.

[41] Scott, J., Social Network Analysis : A Handbook, SAGE, 2000.

[42] Sebastiani, F., "Machine Learning in Automated Text Categorization", *ACM Computing Surveys*, Vol. 34, No. 1, 2002, pp. 1-47.

[43] Sebastiani, F., "Classification of Text, Automatic", The Encyclopedia of Language and Linguistics 14, 2nd edition, Elsevier Science Pub., 2006.

[44] Stanvrianou, A., Andritsos, P., and Nicoloyannis, N., "Overview and Semantic Issues of Text Mining", *ACM SIGMOD Record*, Vol. 36, No. 3, 2007, pp. 23-34.

[45] Wang, W. F., Chung, Y. L., Hus, M. H., and Keh, A. C., "A Personalized Recommender System for the Cosmetic Business", *Expert Systems with Applications*, Vol. 26, No. 3, 2007, pp. 427-434.

[46] Witten, I. H., Text Mining, Practical Handbook of Internet Computing, edited by M. P. Singh, CRC Press, 2004.

[47] http://www.bloter.net/archives/105165, 2012.4.11.

[48] http://www.etnews.com/news/computing/informatization/2634465_1475.html, 2012.8.21.

## 저자소개


**현 윤 진**

현재 국민대학교 비즈니스IT전문대학원에서 비즈니스IT를 전공하고 있다. 국민대학교 비즈니스IT학사 학위를 취득하였으며, 주요 관심분야는 텍스트 마이닝 및 데이터 마이닝이다.

**한 희 준**

현재 한국과학기술정보연구원 NTIS센터에서 선임연구원으로 재직 중이다. 전북대학교 정보통신공학학사 학위를 취득하고 KAIST 전자공학과에서 영상신호처리를 전공하여 석사학위를 취득하였다. 2004년부터 한국과학기술정보연구원에 근무하면서 정보검색엔진 개발, 정보유통 체제 구축과 관련한 업무를 수행하였으며, 현재는 국가 R&D 정보의 원활한 유통체제 구축에 노력하고 있다. 주요 관심분야는 정보검색, 개인화 서비스, 신호처리 등이다.

**최 희 석**

현재 한국과학기술정보연구원 NTIS센터에서 선임연구원으로 재직 중이다. 부산대학교 컴퓨터공학과에서 학사 학위를 취득하고, 동 대학원에서 소프트웨어공학을 전공하여 석사 및 박사학위를 취득하였다. 2004년~2005년까지는 한국전자통신연구원에서 근무하며 센서 네트워크 기술을 연구하였고, 2006년부터 한국과학기술정보연구원에 근무하면서 국가R&D정보관리체계 연구 및 국가과학지식정보서비스(NTIS)의 구축에 노력하고 있다. 주요 관심분야는 R&D정보관리체계, 빅데이터 연계・활용, S/W아키텍처 등이다.

**박 준 형**

안양대학교에서 디지털미디어 공학사를 취득하였으며, 국민대학교 비즈니스IT전문대학원에서 석사학위를 취득하였다. 주요 관심분야는 소셜네트워크분석 및 응용이다.

**이 규 하**

원광대학교 정보전자상거래학과에서 학사 학위를 취득하였으며, 현재 국민대학교 비즈니스IT전문대학원에서 비즈니스IT를 전공하고 있다. 주요 관심분야는 소셜 네트워크분석 및 응용이다.

**곽 기 영**

현재 국민대학교 경영대학 경영정보학부 교수로 재직 중이다. 서울대학교 경영대학을 졸업하고 KAIST 경영과학과 및 테크노경영대학원에서 석사 및 박사학위를 취득하였다. 주요 연구관심분야는 IT-enabled organizational agility, Knowledge management, Social network analysis and its application, System dynamics 등이다.

**김 남 규**

현재 국민대학교 경영대학 경영정보학부에서 부교수로 재직 중이다. 서울대학교 컴퓨터공학과에서 학사 학위를 취득하고, KAIST 테크노경영대학원에서 Database와 MIS를 전공하여 경영공학 석사 및 박사학위를 취득하였다. 한국정보기술응용학회 부회장, 한국경영정보학회 이사, 한국지능정보시스템학회 이사, 한국CRM학회 이사, 한국인터넷정보학회 편집위원, JITAM 편집위원을 역임하였으며, 한국경영정보학회, 한국지능정보시스템학회, 한국정보시스템학회 종신회원 및 한국생산성본부 자문위원으로 활동 중이다. 주요 관심분야는 시맨틱 데이터 관리, 텍스트 마이닝, 데이터 마이닝 등이다.