공학박사 학위논문

기계학습을 이용한
핵심 특허문헌 추출시스템에 관한 연구

Extracting Core Patent Documents
Using Machine Learning

지도교수 류 길 수

2017년 2월

한국해양대학교 대학원
컴 퓨 터 공 학 과
윤 병 수

본 논문을 윤병수의 공학박사 학위논문으로 인준함.

위 원 장 공학박사 김 재 훈 (인)

위 원 공학박사 이 장 세 (인)

위 원 공학박사 하 윤 수 (인)

위 원 공학박사 이 기 욱 (인)

지도교수 공학박사 류 길 수 (인)

2016년 12월 27일

한국해양대학교 대학원

목 차

List of Tables

List of Figures

Extracting Core Patent Documents Using Machine Learning

Yoon, Byung Soo

Department of Computer Engineering,
Graduate School of Korea Maritime and Ocean University

Abstract

The government as well as corporations is promoting research and development (R&D) of new growth engines that are internationally competitive to overcome the ongoing global economic downturn. To do this, they set the direction of the R&D by using some patent trend analysis from the stage of planning and evaluation of the R&D in order to create valuable patented technology with international competitiveness.

Such a patent trend analysis, however, is a time-consuming and error-prone task because it requires patent researchers to manually examine the extracted candidate patent documents one by one and to understand the patent technology out of their expertise. This is a serious problem.

In this dissertation, we propose a method for extracting core patent documents using information retrieval and machine learning. The method contains three steps: 1) extract valid patent documents from retrieved patent documents using a patent search service; 2) classify the valid patent documents into sub-technology categories; 3) finally extract core patent documents from valid patent documents classified by sub-technology categories. The first step ranks retrieved patent documents to obtain valid patent documents for a given queried technology by cosine similarity between the vector of each retrieved patent document and that of the technical summary as the queried technology. The second step classifies valid patent documents into sub-technology categories using a five layered neural network, of which the input is TF-IDF weights and technology-related weights for each valid patent documents. The final step extracts core patent documents from the valid patent document classified by sub-technology categories. In detail, valid patent documents is ranked by linear combination of patent feature values (for instance, impact factor, the number of family nations, cosine similarity, and so on) and a patent feature priority.

For the evaluation, we analyzed patent trends on radiopharmaceuticals as an example. The patent search service retrieved 4,603 candidate patent documents for a technical summary as a queried technology. We compared the results of the proposed system and those obtained manually by a patent investigator in time and accuracy. First, in the execution time, it takes 13,095 minutes to perform manual operations, while the proposed system performed the same operations for 134 minutes. It is 97 times as fast as the manual operations can. And the proposed system have shown the accuracy of 86.88% for extracting valid patent documents, the accuracy of 91.08% for classifying into detailed technology categories, and the accuracy of 75.76% for extracting core patent documents.

Consequentially, we have shown that the proposed system is effective because it helps patent researchers to save the time and to reduce the errors. In the future, we will improve the performance of the proposed system in accuracy using a cutting-edge technology like deep learning and apply to several areas except radiopharmaceuticals.

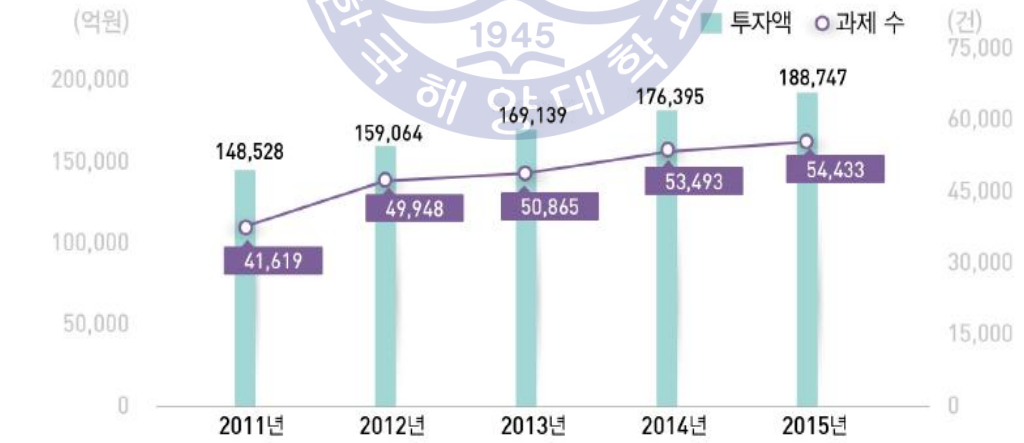
KEY WORDS: Patent Trend Analysis; 특허동향분석; Machine Learning; 기계학습; Cosine Similarity; 코사인 유사도; Neural Network; 신경망;

제 1 장 서 론

1.1 연구의 배경

세계적인 경기침체 등 대내외 경제여건이 지속적으로 악화되면서 실물경제의 위축에 따른 경제위기의 극복방안으로 새로운 신성장동력을 찾기 위하여, 2009년 정부가 주도적으로 연구개발에 나서고 있다[1]. 이러한 경제 성장을 위하여 기술개발과 개발된 기술이 사업화될 수 있도록 정부는 기업, 대학교 및 정부출연 연구기관을 중심으로 연구개발을 활발하게 지원하고 있다.

이에 우리나라 정부도 그림 1.1에서와 같이, 국가연구개발사업의 투자액을 지난 5년 동안 평균 약 6.2%로 증액하여 지원하고 있다. 그에 따라 연구개발의 과제수도 역시 평균 약 7.2% 신장되었다[2, 3].



자료) 미래창조과학부·한국과학기술기획평가원, 「2015년도 국가연구개발사업 조사·분석 보고서」, 2016. 7

그림 1.1 최근 5년간 국가연구개발사업의 투자액 및 과제수

Fig. 1.1 The amount of investment and the number of projects in national research and development for the last 5 years

이러한 연구개발 기술에 대하여 인터넷 등의 정보통신망의 발전과 다국적기업에 의한 세계화의 영향으로 국제적인 경쟁력이 요구된다. 따라서 기술의 국제적인 경쟁력 강화를 위하여 개발된 기술과 그 기술이 반영된 제품의 세계적인 시장 확보를 위한 기술의 분석 및 시장조사의 필요성이 높아지고 있다. 그리고 집약된 기술에 대한 보호와 연구개발 기술의 국제적인 경쟁력을 제공할 수 있는 지적재산권에 대한 사회적 관심도 역시 높아지고 있다. 기술과 관련된 지적재산권으로서는 특허와 실용신안(이하, 특허)이 있다. 이러한 특허는 국제주의를 취하고 있어 세계적으로 새로운 기술의 여부를 확인할 수 있으며, 특허의 공개제도를 통해 국제적으로 기술이 공개되어 동일하거나 유사한 기술에 대하여 보호를 받을 수 있다.

따라서 정부는 연구기관의 연구개발의 기획 및 평가단계에서부터 세계적인 기술성 및 시장성의 분석 및 조사에 특허정보를 적극적으로 활용하여 연구개발의 방향성을 제시하고 있다. 그리고 기술의 연구개발에 대한 중복지원을 방지하기 위하여, 정부는 2004년 국가연구개발사업 효율화를 위한 특허정보 활용확산 계획 및 2005년 시행된 국가연구개발사업의 관리 등에 관한 규정을 제정하였다[4, 5].

기업들도 지식재산권을 기반으로 한 경영전략을 수립하기 위하여, 연구개발에 있어 기술개발의 방향성을 제시하고, 향후 성장가능성과 시장성이 높은 기술을 확보할 필요성을 느끼고 있다. 그리고 현재 보유중인 특허들을 기반으로 확장된 기술을 연구하거나 사업화를 위해, 관련기술에 대하여 국내뿐만 아니라 국제적인 특허기술의 동향에 대한 관심이 높다.

이에 정부와 기업에서 시행되는 연구개발 및 연구기획에 있어, [6]의 연구에서는 미진한 공백기술 영역을 파악하는 특허동향분석을 통해 기술의 개발방향 및 주요 기술에 대한 전망을 예측하고자 하였다. 그리고 [7]의 연구에서는 주제기반 특허분석을 통해 기술예측 시스템에 대한 개발을 수행하였다.

그리고 이미 개발 중인 기술이 종래 특허기술의 권리범위에 속하지 않도록 관련 특허에 대한 분석을 통해 권리범위를 회피할 수 있는 방안을 사전에 파악하여 기술을 개발할 수 있도록 특허동향분석이 많이 활용된다. 그리고 정부와 기업은 연구개발 과제를 선정하는 단계에서도 해당분야의 선행특허의 존재여부 등을 미리 조사하여 과제 선정에 활용함으로써 향후 개발된 기술에 대한 특허권의 확보방안도 마련하고 있다[8].

정부와 기업은 이러한 기술의 연구개발에 앞서 지식재산권의 전문가인 특허조사원에게 이미 공개되어 있는 연구개발 기술과 관련된 특허기술에 대한 기술의 분석 및 시장성 조사를 수행하는 특허동향분석을 의뢰한다. 특허조사원은 정부나 기업으로부터 특허동향분석에 대한 요청이 오면, 그림 1.2와 같이, 수작업으로 특허동향분석을 수행한다.

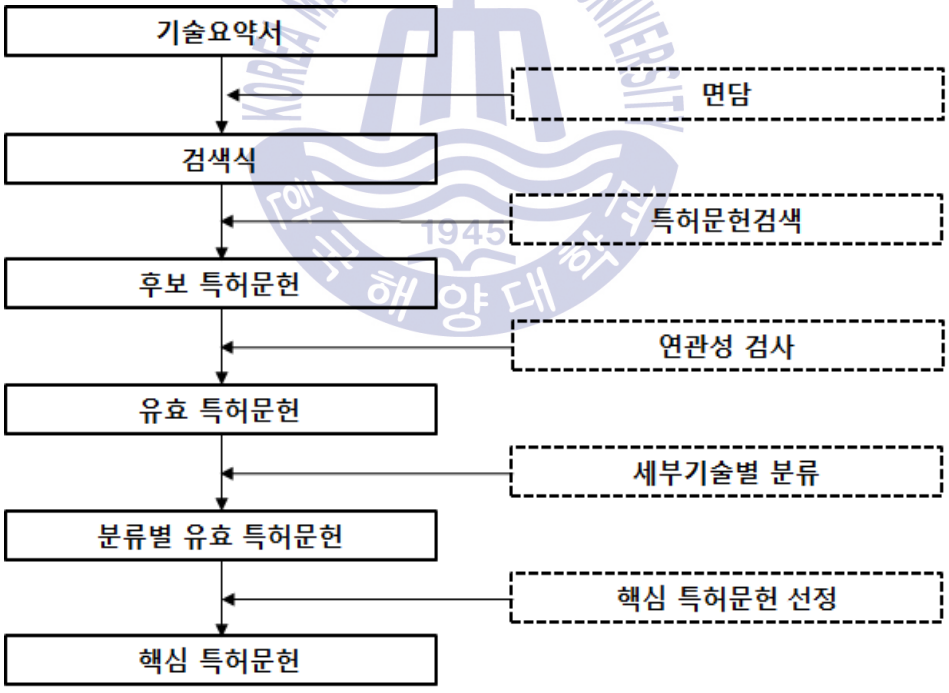


그림 1.2 특허동향분석의 절차

Fig. 1.2 The procedure of patent trend analysis

특허동향분석은 먼저 특허조사원이 기술요약서 및 연구기관의 담당자와 면담을 통해 연구개발 기술을 파악한다. 파악된 연구개발 기술의 구체적인 구성요소를 기반으로 특허조사원은 색인어를 도출하여 검색식을 작성한다. 특허조사원은 작성된 검색식을 이용하여 외부의 특허검색서비스를 이용하여 특허문헌들을 검색한다.

특허조사원은 특허검색을 통하여 외부의 특허검색서비스에서 제공하는 후보 특허문헌 목록을 추출하여 파일로 다운받는다. 다운받은 후보 특허문헌 목록은 색인어를 포함하는 모든 특허문헌이 검색되기 때문에 찾고자 하는 연구개발기술과 관련성이 없거나 관련성이 낮은 특허문헌을 포함하고 있다. 따라서 특허조사원은 추출된 후보 특허문헌 목록에 대하여 기술의 연관성을 검사하여 연관성이 부족한 특허문헌을 제거한 유효 특허문헌만을 추출한다.

추출된 전체 유효 특허문헌을 이용하여 특허조사원은 여러 가지 특허기술과 관련된 동향지표를 생성한 후 전체 특허동향에 대한 정량분석을 수행한다. 다음으로 특허조사원은 의뢰된 연구개발 기술을 상위개념으로 하여 세부기술 분야를 나누고, 나눈 세부기술별로 유효 특허문헌을 분류한다. 분류된 유효 특허문헌을 기반으로 세부기술별 특허동향지표를 작성하여, 기술분야별 정량분석을 수행한다.

마지막으로, 특허조사원은 세부기술별 유효 특허문헌 중에서 개발하고자 하는 기술과 가장 유사한 특허문헌을 핵심 특허문헌으로 추출한다. 추출된 핵심 특허문헌에 기재되어 있는 특허기술에 대한 기술범위 및 권리범위를 분석하여 연구개발 기술과 비교하는 정성분석을 수행한다. 이러한 분석된 결과를 토대로 특허조사원은 특허동향분석 보고서를 작성하게 된다.

특허동향분석에 필요한 특허문헌의 검색은 기술요약서로부터 검출한 색인어를 이용하는 키워드 기반의 불리언(boolean) 검색으로 이루어지고 있다. 그리고 대부분의 특허검색서비스도 불리언 검색을 주로 지원하고 있다. 하지만 일반문서에 비해 특허문헌을 기재하는 명세서는 일정한 형식의 여러 항목이 존재하고, 그 형식에 맞춰 특허문헌을 기재하고 있는 변리사 및 명세사의 연구개발기술에 대한 단어의 선택이나 표현이 너무 다

양하고, 기술적 표현 또는 법적 표현이 많은 특징으로 정확한 특허문헌의 검색이 어렵다. 그리고 기존의 특허검색서비스에 있는 특허문헌 데이터베이스가 가지고 있는 구조상 한계로 인해 최근에 많이 연구되고 있는 문장검색의 정확도에 대한 신뢰성이 떨어져 있다.

이와 같은 이유로 인해 특허조사원이 검색식을 작성함에 있어서도, 연구개발 기술에 대한 색인어를 이용하여 1차적으로 작성된 검색식으로 특허검색서비스에서 검색하여 추출한 후보 특허문헌의 내용, 특허문헌의 수량, 검색 주제에 맞는 국제특허분류(International Patent Classification, 이하 IPC)코드 등을 확인하여 최적의 결과를 얻기까지 검색식을 수차례 수정하여 최종 검색식을 결정하므로 상당한 시간이 소요된다. 또한, 특허조사원은 결정된 최종 검색식으로 특허검색서비스에서 추출한 후보 특허문헌에 기재된 기술과 연구개발 대상 기술 간의 동일 또는 유사여부를 판단하여 유효 특허문헌을 추출한다. 그런데 이러한 유효 특허문헌을 추출하는데 있어, 첫째, 검색된 후보 특허문헌의 기술내용을 하나씩 판단하여야 하는데 최근 지식재산권에 대한 관심 및 필요성의 증가로 인해 매년 출원되는 특허문헌의 양이 급격히 늘어나 판단해야할 문헌의 수가 점점 많아지고 있다. 둘째, 소비자의 기술수요에 따른 새로운 기술분야가 많이 나타나고 있으며, 셋째, 다른 기술분야를 응용하거나 2개 이상의 기술분야들을 융합하는 융합기술의 출현으로 하나의 특허기술의 내용을 파악하는데 복수영역의 기술을 이해해야 하는 현실적인 어려움을 겪고 있다.

특허조사원이 추출한 유효 특허문헌에 대하여 세부기술별로 분류하는데 있어서도 추출된 유효 특허문헌을 대상으로 기술이나 기술의 적용 범위를 파악하여 분류하여야 한다. 그리고 연구개발 기술과 관련된 유효 특허문헌에 따라 기술을 분류하여야 하므로 기술의 분류 기준이 명확하지 않으며, 명확한 선정 기준이 되는 항목들을 선택하였음에도 이를 정확히 반영하기 위해서는 파악해야할 특허문헌이 많다. 이에 반해 특허동향분석을 의뢰하는 정부나 기업에서는 비교적 빠른 분석결과를 요구하여, 시간적으로나 정확성에 있어 한계를 보이고 있다.

1.2 연구의 목적 및 범위

본 논문에서는 이러한 특허동향분석에 있어 발생하는 시간적인 한계를 해결하고자 핵심 특허문헌 추출시스템을 제안한다. 제안된 추출시스템은 특허조사원에 의해 수작업으로 이루어져 시간이 많이 요구되는 그림 1.2의 특허동향분석의 절차 중에서 연관성 검사, 세부기술별 분류 및 핵심 특허문헌의 선정을 자동적으로 수행한다.

연관성 검사와 관련하여 기존의 특허조사원이 특허검색서비스에서 추출된 후보 특허문헌 각각에 대하여 연구개발기술과 관련성을 파악하였다. 이를 추출시스템에서는 기술정보인 기술요약서의 문서벡터와 각 후보 특허문헌의 문서벡터를 연산하여, 이들 문서벡터간의 유사도(similarity)를 측정하여 파악한다.

세부기술별로 분류에서는 특허조사원이 추출된 유효 특허문헌에 대하여 기술의 내용을 파악하여 세부기술별로 유효 특허문헌을 분류하였다. 이러한 분류작업을 추출시스템은 각 유효 특허문헌의 색인후보어에 대한 TF-IDF(Term Frequency-Inverse Document Frequency)가중치와 분류 관련 색인후보어에 대한 분류가중치를 신경망(neural network) 알고리즘의 입력파라미터로 이용하여 분류확률을 순위화하고, 분류확률이 상대적으로 높은 세부기술을 선택함으로서 유효 특허문헌을 기술별로 분류한다.

핵심 특허문헌의 선정은 특허조사원이 분류된 유효 특허문헌에 기재되어 있는 기술내용과 기술에 대한 평가 등을 참조하여 세부기술별로 핵심 특허문헌을 선정하였다. 추출시스템에서는 이러한 선정을 유효 특허문헌에 포함된 항목들 중에서 기술의 평가와 관련된 항목과 유사도에 대하여 우선순위를 부가하여 선정된 항목의 값과 우선순위가중치를 선형 연산하여 구한 선정값을 순위화하여 지정된 수의 핵심 특허문헌을 선정한다.

핵심 특허문헌 추출시스템의 평가와 관련하여 추출된 유효 특허문헌, 분류된 세부기술별 유효 특허문헌 및 선정된 핵심 특허문헌을 각각 특허조사원이 수작업을 통하여 얻어 결과와 비교하여 정확성을 확인한다. 그리고 특허조사원이 각 작업을 통하여 수행하는데 요구되는 시간과 추출시스템을 통해 실제 수행한 시간을 각각 비교하고자 한다.

1.3 연구방법

본 논문에서는 연구개발기술과 관련된 기술요약서를 이용하여 종래의 특허조사원에 의해 수작업으로 수행하였던 특허동향분석에 필요한 유효 특허문헌의 추출, 세부기술별 분류 및 핵심 특허문헌을 선정하기 위한 핵심 특허문헌 추출시스템을 설계 및 구현하고자 한다. 이를 위해 기존의 일반 문서를 대상으로 수행하였던 정보검색 및 기계학습의 알고리즘 중에서 TF-IDF가중치, 유사도 측정 알고리즘, 신경망 알고리즘을 특허문헌에 응용하고자 한다.

유효 특허문헌의 추출과 관련해서는 기술요약서와 각 후보 특허문헌의 연관성을 판단하기 위하여 유사도 측정 알고리즘을 이용한다. 문서간의 유사도 측정을 위해 기술요약서의 문서벡터와 각 후보 특허문헌의 문서벡터를 측정한다. 기술요약서에 대한 문서벡터는 기술요약서에 기재된 색인후보어에 대한 TF-IDF가중치를 구하고, 각 후보 특허문헌의 문서벡터는 색인후보어의 TF-IDF가중치뿐 아니라 항목별가중치 및 비색인어가중치를 이용하여 구한다.

항목별가중치는 여섯 세트로 구성하여 각 세트별로 적용한 유효 특허문헌의 결과와 기존의 특허조사원이 수행한 결과를 비교하여 정확률을 측정하고, 높은 정확률을 제공하는 항목별가중치를 확인한다. 그리고 추출시스템의 수행시간도 수작업의 수행시간과 비교하여 추출시스템의 성능을 확인한다.

세부기술별 분류에 있어서도 다섯 계층의 신경망 알고리즘을 이용하여 추출된 유효 특허문헌의 색인후보어 TF-IDF가중치와 분류가중치를 입력데이터로 이용하여 분류확률을 순위화한다. 순위화된 분류확률 중에서 상대적으로 높은 확률값을 이용하여 분류한다. 추출시스템을 이용하여 분류한 결과와 특허조사원이 분류한 결과를 비교하여 정확률을 비교하여 다른 분류 알고리즘과의 분류성능을 확인하고, 수행시간을 비교하여 성능을 평가하고자 한다.

마지막으로 핵심 특허문헌의 선정은 분류된 유효 특허문헌의 항목 중에서 유사도와 기술평가와 관련된 항목을 선정한다. 그리고 선정된 항목에 대한 우선순위를 평가하여 항목값과 우선순위에 따른 가중치를 입력데이터로 선정값을 구하여 순위화하여 지정된 수의 핵심 특허문헌을 선정한다. 그리고 넷 세트의 우선순위가중치를 이용하여 각 세트별로 선정값을 구하여 핵심 특허문헌을 선정한 결과와 특허조사원이 선정한 핵심 특허문헌을 비교하여 정확률을 확인하고, 우수한 결과를 보이는 우선순위가중치를 선별한다. 그리고 추출시스템의 수행시간을 측정하여 특허조사원의 수행시간과 비교함으로써, 추출시스템의 성능을 평가한다.

본 논문의 구성은 다음과 같다. 2장에서는 본 논문과 관련하여 정보검색 및 기계학습과 관련된 연구를 확인하고, 3장에서는 본 논문에서 제안하고 있는 핵심 특허문헌 추출시스템의 구성과 이를 이용한 알고리즘에 관하여 다룬다. 4장에서는 핵심 특허문헌 추출시스템을 이용한 실험 및 평가를 통해 시스템의 성능을 확인하며, 5장에서는 결론 및 향후 과제에 대하여 논한다.

제 2 장 관련 연구

2.1 정보검색

2.1.1 색인어 검출 및 가중치

형태소 분석이란 문장에 기재되는 단어를 구성하는 각각의 형태소들을 인식하고 불규칙 활용이나 축약, 탈락 현상이 일어난 형태소의 원형을 복원하는 과정을 말한다[9]. 태거는 형태소 분석기가 분석한 결과 중에서 전체 문장의 문맥에 적합한 하나의 분석결과를 선택하여 품사를 태깅하는 모듈을 말한다[10].

형태소 분석기는 먼저 입력된 특정 문서에 대하여 대상 문장 추출, 문장부호, 숫자, 특수문자열 처리 및 어절 분리와 같은 전처리과정을 수행한 후, 단어의 다양한 변형규칙에 대하여 원형 복원 규칙을 적용하여 분석후보를 생성한다. 그리고 생성된 분석후보를 대상으로 다양한 어휘 사전 정보 및 각종 제약 조건을 반영하여 의미를 가지는 최소 단위인 형태소를 생성한다. 생성된 형태소에 대하여 단어의 중의성 해소, 복합어 추정, 조사 생략, 준말 처리 등의 후처리 과정을 수행한다[11].

특정문서로부터 특정 품사의 단어를 출하기 위해서는 이러한 형태소 분석기를 이용하여 단어를 검출한다. 특허문헌을 검색하는데 있어서도 특허검색서비스에서 검색식을 활용하여 검색하는데, 기술에 대한 문서로부터 형태소 분석기를 활용하여 검색식을 구성하는 색인어에 사용할 명사를 검출할 수 있다.

TF-IDF가중치는 문서를 구성하는 단어의 벡터값을 기반으로 하는 벡터공간모델로서, 정보검색(information retrieval), 주제탐색(topic detection), 문서요약(document summarization), 문서유사도(document similarity), 문서분류(text categorization) 등 텍스트를 이용한 데이터마이닝(data mining), 기계학습(machin learning) 분야에서 많이 사용하는 가중치이다[12]. 그리고 TF-IDF가중치는 복수의 문서가 있을 때 특정 단어가 전

체 문서에서 중요도가 얼마인지 여부에 대하여 판단하기 위해 나타내는 통계적 수치이다. 이러한 TF-IDF가중치는 문서 등에서 색인어 검출, 검색 엔진에서 검색 결과의 순위 결정, 문서들 간의 유사도 산출 등의 용도로 사용된다.

TF(term frequency)값은 특정 단어가 하나의 문서 내에서 얼마나 자주 사용되는지, 즉 특정 단어의 빈도를 표시하는 값으로서, 특정 단어가 문서에서 사용되는 빈도가 높은 단어일수록 TF값이 높을 것이며, 높은 TF값을 가진다는 것은 해당 문서에서 중요한 단어라고 볼 수 있다[13]. 이러한 단어의 TF값은 식 (2.1)과 같이, 일반적으로 문서 내부의 단어 출현 빈도수를 모든 단어의 총 출현 빈도수로 나누어 정규화하여 사용하는 경우가 많다. 여기서, $tf_{i,j}$는 문서 d_j에 대한 특정 단어 t_i의 출현율이다. 그리고 $f_{i,j}$는 문서 d_j에서 특정 단어 t_i의 출현 빈도수이며, $\sum_{i=1}^{n} f_{i,j}$는 문서 d_j에 있는 모든 단어의 수를 나타낸다.

$$tf_{i,j} = \frac{f_{i,j}}{\sum_{i=1}^{n} f_{i,j}} \tag{2.1}$$

DF(Document Frequency)값은 특정 단어 자체가 문서들의 집합에서 반복적으로 사용되는지를 나타내는 값이다. 이러한 DF값의 역수인 IDF(Inverse Document Frequence)값은 식 (2.2)와 같이, 대상이 되는 문서집합의 전체수에 특정 단어가 나타난 문서의 수로 나눈 값이다.

$$idf_i = \log \frac{|D|}{|\{d_j \mid t_i \in d_j\}|} \tag{2.2}$$

여기서, idf_i는 전체 문서집합에 대한 단어 t_i의 기여율이다. 그리고 $|D|$는 문서집합의 총 개수이고, $|\{d_j \mid t_i \in d_j\}|$는 특정 단어 t_i를 포함하는 문서 d_j의 개수이다. 그리고 여기서 분모인 특정 단어 t_i를 포함하는 문서의 수가 매우 작은 경우, idf_i가 매우 커질 수 있는 문제를 방지하기 위하여 로그(log)를 취한다.

TF-IDF가중치는 식 (2.3)과 같이, TF값과 IDF값을 곱한 값으로서 특정 단어가 하나의 문서에서는 빈도수가 높고, 전체 문서집합을 기준으로는 해당 단어가 출현하는 문서들의 수가 적은 단어가 중요한 단어로 평가한다[14]. TF-IDF가중치를 통한 특허 문서 검색의 기본 아이디어는 각 단어의 가중치를 계산 후, 상위 적당 개수의 가중치를 가지는 색인어를 선정하는 것이다.

$$tfidf_{i,j} = tf_{i,j} \times idf_i \tag{2.3}$$

TF-IDF가중치는 내용기반의 문서추천에 널리 활용되고 있다. [15]의 연구에서는 특정한 단어가 문서에서 얼마나 중요한지를 벡터공간(vector space)에 상대적인 가중치로 표시하는 방법으로, 특허의 핵심적인 내용을 파악하는데 활용되었다.

[16]의 연구에서는 특허정보 검색을 위하여 특허문서의 특징을 기반으로 한 TF-IDF가중치를 이용한 벡터 스페이스 검색 모델이 제안된다. 특허문헌은 텍스트 기반의 문서이기 때문에 특허문헌의 각각의 단어에 더 정확한 가중치 부여를 위한 텍스트 필드의 색인어에 TF-IDF가중치를 부여한 벡터로 표현하였다.

특허 문헌을 분석함에 있어, 불용어(stopword)는 특허맵(patent map)의 매핑을 지연시키는 요소이다. 그래서 [17]의 연구에서는 TF-IDF가중치가 특허맵에서 검색된 단어의 불용어를 걸러주는 역할을 할 수 있어 TF-IDF가중치가 특허맵을 매핑하는데 매우 중요하고 실용적인 알고리즘임을 확인할 수 있었다.

[7]의 연구에서는 미래의 기술을 예측하는 시스템 개발에 있어, 주제기반의 특허분석에 있어 특허문헌에 대하여 기술군의 집합을 형성하기 위해 TF-IDF가중치를 이용하여 출현단어의 중요도를 측정하였다. 텍스트 문서 분류에 있어서도 문서의 특징들에 가중치를 부여하여 벡터화하는 문제도 연구되고 있는데, [18]의 연구에서는 Reuters-21578을 SVM(Support Vector Machine) 분류기를 이용하여 문서 범주화 실험을 수행하기 위해 자질 값을 생성하는데 있어서, IDF값, DF값, TF값 및 이진(Boolean) 등은 실제 분류 성능에는 그다지 큰 차이를 보이지 않았다.

[19]의 연구에서는 기술정보 문서를 디렉터리 기반으로 분류하기 위하여 SVM 분류기에 문서 정보를 학습시키는데, 문서의 텍스트 데이터를 TF-IDF가중치로 사용하여 벡터 공간 모델로 구성하였다. 실험에 사용된 문서 기술정보 문서는 정보보호센터(CERT-KR)와 정보보호진흥원(KISA)에서 수집한 문서를 사용하였다.

2.1.2 문헌유사도

유사도는 문서와 문서 사이의 어느 정도 관련성이 있는지를 수치적으로 계산하는 것이다. 하나의 문서는 이를 벡터로 표현할 수 있기 때문에 유사도 측정방법 또한 벡터 계산으로 이루어진다[20]. 유사도 측정에 사용되는 유사계수는 대상 문서와 문서간의 유사성이나 상이성을 측정하는 척도로 거리 계수(distance coefficient), 연관 계수(association coefficient), 상관 계수(correlation coefficient), 확률적 유사계수(probabilistic similarity coefficient)가 있다.

거리 계수는 벡터 공간상에서 대상간의 비유사성을 측정하는 거리를 이용하여 방법으로, 유사도가 높은 두 문서간의 거리는 짧다. 이러한 거리 계수를 측정 방법에는 코사인 거리(cosine distance), 유클리디언 거리(Euclidean distance), 민코스키 메트릭스(Mincowski metrics), 시티 블록 거리(city block distance) 등이 있다.

연관 계수는 두 문서에 대한 자질을 나타내고 있는 속성들 간의 일치 정도를 측정 방법으로, 자카드 계수(Jaccard coefficient), 다이스 계수(Dice coefficient) 등이 있다. 또한 상관 계수는 비교하고자 하는 두 문서를 표현하는 속성들의 벡터 쌍에 대한 독립성을 측정하는 방법으로서, 대표적으로 피어슨 상관계수(Pearson correlation coefficient)가 있다.

확률적 유사계수는 정보량에 관한 공식을 근거로 두 사건의 확률 변수간의 의존 관계를 정량적으로 나타낸 것이다. 새넌(Shannon)의 정보이론에 기초한 상호정보량(mutual information)이 확률적 유사계수에 속한다[21].

단어를 기반으로 하는 문서간의 유사도를 판단에 있어서, 일반적으로 문서가 단어 벡터화 되어있을 때, 학습 모델의 이해도가 쉬우며, 단어 벡터 간 상관관계를 수량화되어 단어의 가중치와 같은 숫자데이터를 입력데이터로 하는 거리 계수를 이용한 유사도 측정방법이 많이 사용되어 진다[22]. 이러한 거리 계수는 데이터 간 유사도 측정의 주안점으로 데이터 특질 및 그 경향 혹은 방향성이 중요시 되는 분야에서 폭넓게 사용된다[23].

거리계수 중에서 많이 사용되는 코사인 거리는 두 문서 벡터 간의 각에 대한 코사인 값을 이용하여 유사도를 측정하는 것으로, 식 (2.4)로 표현된다[24]. 여기서 d_i벡터는 $w_{i,1}, w_{i,2}, \cdots, w_{1,n}$로 이루어지고, d_j벡터는 $w_{j,1}, w_{j,2}, \cdots, w_{j,n}$이다. 코사인 거리를 이용한 유사도 값은 0 ~ 1사이의 값의 가지게 되어 확률값으로도 사용될 수 있다.

$$sim(d_i, d_j) = \cos\theta = \frac{d_i \cdot d_j}{|d_i| \cdot |d_j|} = \frac{\sum_{k=1}^{n}(w_{i,k} \times w_{j,k})}{\sqrt{\sum_{k=1}^{n}(w_{i,k})^2} \times \sqrt{\sum_{k=1}^{n}(w_{j,k})^2}} \quad (2.4)$$

문서 벡터의 유사도를 이용하여 문서내용의 유사성을 판단하는 연구가 활발히 이루어지고 있다. [25, 26]의 연구에서는 기계학습을 이용한 문서 간의 유사도를 결정하는 연구와 관련하여 유사도 측정 함수의 종류, 적용되는 분야 등을 벡터 정보 개체와 함수들의 대수적 특성에 따라 비교 분석하였다.

[27]의 연구에서는 용어의 유사도와 관련하여 특정 임계값을 설정하고, 임계값보다 큰 유사도를 가지는 용어들만을 사용하는 경우에서 여러 실험에서 결과가 오히려 나빠지는 것을 확인하였다. 그래서 [28]의 연구에서는 이를 보완하기 위해 단어를 개념 공간상에 하나의 개념으로 표현한 후, 문서에 대한 단어의 가중치를 반영하여 단어 간 유사도를 계산한 확률 기반의 질의 확장으로 수행하는 경우 좋은 검색 성능을 보였다.

[29]의 연구에서는 비감독 학습방법에 있어서도, 문서 특징들의 유사도를 이용하여 비감독(unsupervised) 특징 선택에 있어 효율성을 향상시켰다. 또한 [30]의 연구에서는 문서 분류에 있어 문서의 품사 중 명사를 중심으로 용어를 검출하고, 이와 같은 검출된 특징들을 클러스터링을 기반으로 나이브 베이지안(naïve Bayesian), 결정트리(decision tree) 및 SVM 분류기들을 이용하여 비교 실험하였다. 그리고 [31]의 연구에서는 WebKB, Reuters 및 RCV1 등의 문서 분류 및 클러스터링을 수행하는데 있어서도 이들 문서들의 유사도 측정 방법을 이용하여 수행하였다. 한편 특허와 관련한 문헌의 경우에는 특허문서 내용 이외에도 여러 속성 정보들이 존재한다.

[32]의 연구에서는 특허들 간의 참조를 통한 유사 특허에 대한 네트워크도 존재하지만, 특허 문헌에 기입된 출원인의 정보 등의 속성으로 연결되는 네트워크도 역시 존재한다. 이러한 관점으로 특허 문서를 네트워크로 표현하여 데이터를 찾는 연구도 진행되고 있다. 보다 진보된 방안으로, [33]의 연구에서는 특허가 포함된 정보와 특허내의 존재하는 레퍼런스 관계를 이종의 네트워크로 이용한 Patentline방법을 연구하였는데, 이는 특허를 그래프로 표현하고, 정보와 정보를 지배하고 있는 특허를 추출하여 추출된 핵심 특허 사이의 경로를 통해 특허 기술의 흐름 등을 찾는 연구도 진행 되었다. 하지만 이러한 이종 네트워크로 표현하는 방법이 아닌 특허의 레퍼런스 정보만으로 네트워크로 표현하는 방법에서 국제 특허의 경우에는 레퍼런스 정보를 특허에 포함하고 있기 때문에 적용이 가능하지만, 국내 특허의 경우에는 맞지 않는 부분이 존재하는 문제가 있다. 이에 [34]의 연구에서는 특허의 유사도를 이용한 검색에 있어 특허의 이종 네트워크를 사용하는 방법이 아닌 네트워크 구조를 통해 숨겨져 있는 특허를 찾도록 하는 유사도를 구현하였다. 그리고 [35]의 연구에서는 문서의 특징에 대하여 차원의 개념을 적용하여 높은 차원의 문서를 낮은 차원의 핑거프린트(fingerprint)로 차원 감소(dimensionality reduction)하여 이를 문서의 특징으로 사용하는 대표적인 모델이 유사도 분석 모델이라고 하였다.

하지만 이는 자연어의 중의성을 반영하지 못하며, 나아가 문법의 표준을 준수하지 않은 문장에 대하여 비교가 어렵다는 단점이 있다. 또한, 유사 문서의 분석은 품질 뿐만 아니라 성능도 매우 중요한데, 문서 n개를 핑거프린트로 비교할 경우, 전수 검사가 필요하며, 이들의 유사도를 구하는 일은 n^2정도의 복잡도가 요구되어 문제가 있다. 이러한 문제를 해결하고자 [36]의 연구에서는 다중 레벨 인덱싱 구조를 이용하여 문서의 줄수와 k비트수를 조정하여 비교 대상문헌간의 유사도를 판단한다. 이런 경우, 동일한 핑거프린트 군집의 수가 줄어들기 때문에 보다 향상된 속도로 비교가 가능해진다.

[37]의 연구에서는 문서에 등장하는 단어를 나열하고, 이 중 일부를 문서의 특징으로 추출하는 기법을 제안하였다. 그리고 [38]의 연구에서는 현실적으로 가장 널리 쓰이는 유사도 분석 모델이지만, 키워드를 식별하기 위해 전문가의 휴리스틱한 판단이 필요하다는 단점이 있다고 발표하였다.

[39]의 연구에서는 정부 또는 기업의 연구개발 과제 측면에서 유사도를 측정하기 위하여 포괄성형망(comprehensive star network) 알고리즘을 이용하였다. 이는 종래 키워드 기반 유사도 분석 모델의 한계점을 개선하기 위해 고안된 모델로, 연구개발 사업을 구성하고 있는 과제들의 과학기술표준분류항목을 추출하여, 각 사업별 기술분류에 대한 고유 벡터를 생성하고, 각 사업을 구성하는 과제들의 고유한 기술 패턴을 지정함으로써 이들의 유사성을 포괄성형망으로 표현한다.

[40]의 연구에서는 일본어와 영어 특허문서 간에 기계번역을 위하여 문장구조 유사도와 단어 유사도를 이용한 클러스터링 기반의 통계기계번역 방식을 사용하였다. 이는 문장 및 단어 유사도를 측정하는데 각각 커널과 코사인 유사도를 적용하였고, 두 유사도를 적용하여 말뭉치를 분류하는 과정에서는 *k*-means 알고리즘과 유사한 기계학습 기법을 사용하였다.

2.2 문서분류

종래 문서를 자동적으로 분류하기 위한 문서 범주화 모델로 여러 가지 기계학습 알고리즘들이 제안되었다. 대표적인 기계학습 방법으로는 규칙 기반 모델(rule based model)인 결정트리, 통계 기반 모델(statistics based model) 인 신경망 알고리즘, 예제 기반 모델인 *k*-최근접 이웃법(*k*-Nearest Neighbor, 이하 *k*NN)과 LLSF(Linear Least Squares Fit), 확률 기반 모델인 나이브 베이지안 및 SVM 등이 있다[41-44].

이러한 기계학습을 이용하여 텍스트 분류 또는 문서 분류를 위한 알고리즘에 대한 연구도 활발히 진행되고 있다. [45]의 연구에서는 구조화되지 않은 데이터를 분류하기 위하여 계층적인 분류를 사용하여 많은 수의 문서를 분류하였으며, [46]의 연구에서는 문서를 분류하기 위하여 나이브 베이지안을 이용하여 분류기를 제안하였다.

특허문헌들에 대한 기술분류는 미국의 USPC(US Patent Classification), 일본의 FI(File Index)/F-Term 및 유럽의 ECLA(European CLAssification System) 등으로 국가마다 자체적인 분류코드를 사용하고 있거나, 미국과 유럽이 공통으로 특허를 분류하기 위한 CPC(Cooperative Patent Classification) 등이 있다[47-50]. 이렇게 국가 별로 존재하는 기술분류에 대하여 세계지적재산권기구(World Intellectual Property Organization, WIPO)에서 국제적으로 통일된 특허분류 체계를 수립하고자 협의를 통해 5년 마다 개정되는 IPC코드체계를 확정하고, 그에 따라 분류 대상의 특허 문헌이 어느 기술분류에 속하는지 분류하고 있다[51].

따라서 현재 국내에서도 특허청 산하의 한국특허정보원(KIPI)의 분류 전담 인원이 수동으로 특허청에 출원되는 모든 특허문헌에 대해서 IPC코드를 기반으로 분류하고 있다[52]. IPC코드는 특허문서에서 특허 검색, 선행기술 조사 및 특허맵 작성 등에 매우 유용하게 사용되고 있다[53, 54]. 이렇게 수동적인 분류작업에 따른 시간과 인력의 소모를 감소하고자 기계학습을 이용하여 특허문헌에 대하여 IPC코드를 자동적으로 부여하는 연구가 국내외에서 활발하게 이루어져 왔다.

국가별로도 자국 내의 특허분류체계에 대한 관심도 증대되어 연구가 진행되었는데, [55]의 연구에서는 미국의 국내 특허분류체계인 미국특허분류(US Patent Classification : UPC)의 검색과 자동분류를 위하여 특허문서에서 사용된 용어들을 벡터화하고, 이를 *k*NN 알고리즘을 이용하여 자동 검색 및 분류시스템을 구현하였다. 그리고 [56]의 연구에서는 미국특허를 대상으로 신경망 알고리즘 중 하나로 각 카테고리에 대한 가중치의 벡터를 계산하는 지도 학습 알고리즘인 Winnow 알고리즘을 이용하여 68%의 정확도 성능을 보였다.

[57]의 연구에서는 미국의 UPC 자동분류를 위하여 MI특징선택 방법에 SVM을 결합한 방법을 사용하였으며, 특히 BOW(Bag-Of-Words)와 같은 자연어 기반의 방법을 이용하여 용어들을 처리하였다. 그리고 [58]의 연구에서는 유럽특허청(Euro Patent Office : EPO)에서 발행되는 유럽특허문헌에 대하여 자동적으로 기술분류를 수행하기 위하여 자연어처리기술을 이용하였다.

[59]의 연구에서는 특허문헌을 분류하기 위한 다양한 알고리즘에 대한 성능 비교에 관한 연구도 다양하게 진행되었는데, 알고리즘 선택을 위해 전체 클래스 수준에서 *k*NN(46%), 나이브 베이지안(41%) 및 SVM(48%), SNOW(43%) 등을 이용한 분류기의 정확도 성능을 비교 실험하였다. 또한 [60]의 연구에서는 이 방법을 이용하여 세계지적재산권기구의 CLAIMS(Classification Automated Information System)시스템에 적용 시도를 하였다. 그리고 [61]의 연구에서는 특허문헌의 IPC코드 기반의 자동분류를 위하여 특징추출 방법을 이용하여 SVM(66%), 나이브 베이지안(64%), *k*NN(61%) 및 결정트리(66.7%) 방법의 성능을 비교 실험하였다. 또한, [62]의 연구에서는 특허문서 분류에 있어서, 이러한 각 분류기들의 장단점을 비교하는 실험을 실시하였다.

특허분류의 정확도 향상을 위하여 [63]의 연구에서는 *k*NN, LLSF, 신경망 및 Winnow 알고리즘을 복합적으로 이용하였고, 특허문서 중에서 제목, 요약과 청구항에서 먼저 나오는 200단어, 상세한 설명에서 400단어를 사용하여 기존의 분류기보다 특허문서 분류 정확도가 6.51%가 향상된 분류기에 대한 연구를 수행하였다.

최근 특허출원의 급격한 상승으로 인하여 특허문헌의 양이 증가하면서 기계학습에 있어 학습 데이터의 크기가 점점 더 증가하는 문제가 발생하면서, 시간 복잡도와 자질 공간 증가를 해결하기 위한 연구가 많이 진행되었는데, 그 방법으로서 학습 데이터 문서의 길이를 제한하거나 학습 문서 집합을 샘플링하는 방법 또는 문서의 자질의 선택 방법 등을 제시하고 있다. 특히, 인스턴스 기반 알고리즘인 *k*NN을 적용한 연구가 많았다.

[64]의 연구에서는 코사인 유사도와 유클리드 거리 기법을 적용한 *k*NN 알고리즘을 이용하여 NTCIR-7 patent mining Task를 위한 IPC코드를 분류하는데 시간상의 향상을 보였다. 그리고 [65]의 연구에서는 *k*NN과 재순위 모델을 적용한 IPC코드 분류의 서브그룹 수준의 분류에서 48.86%의 성능을 보였다.

특허문헌에 대한 분류 시스템을 위해 기존의 텍스트 분류기술인 SVM, 나이브 베이시안, 신경망 알고리즘 및 *k*NN 등이 사용되고 있다. 그러나 [66]의 연구에서는 최근에 많이 사용되고 있는 SVM에서 중요한 최적의 커널의 종류와 매개변수를 찾을 수 있도록 HGA(Hybrid Genetic Algorithm)에 관한 연구를 수행하였다.

[67]의 연구에서는 특허문헌의 분류코드인 IPC코드 분류에서 하위 서브 클래스와 서브 그룹 수준의 특허분류 실험에 대하여 *k*NN 알고리즘을 이용하여 40.28%의 적합 문헌 평균 정확률(MAP)을 보이는 결과가 나왔다. 그리고 [68]의 연구에서는 특허문서의 IPC코드 계층 분류를 위하여 중요도가 높지만 출현 빈도의 수가 비교적 적은 용어들을 이용하여 자동분류를 실험하였다.

[69]의 연구에서는 특허 문헌의 국제 분류 체계인 IPC코드 자동분류단계를 하위 서브그룹 수준에서 실험을 하였다. [70]의 연구에서는 IPC코드 자동분류의 서브그룹 단계의 분류 성능을 향상시키기 위하여, 3단계로 구분하여 분류기를 상이하게 진행하여 1단계와 2단계에서는 SVM 분류기를 이용하고, 3단계에서는 *k*NN 분류기를 사용하였다.

기계학습을 이용하여 특허 문헌에 대한 기술분류인 IPC코드를 자동으로 분류하는 연구가 진행되었는데, [71]의 연구에서는 일정기간의 학습 데이터 문헌과 검증 데이터 문헌에 대하여 8개의 기술분야를 대상으로 TF-IDF가중치와 DF가중치 및 이중가중치를 이용한 나이브 베이지안과 SVM의 성능을 비교함에 있어 기술분야에 따라 정확도의 차이가 있었으나, 평균적으로 90.3%의 정확도를 보였다. 그러나 [72]의 연구에서는 특허문헌에 기재된 단어와 단어사이의 거리 즉 친밀도를 이용하여 클러스터링한 후, IPC코드의 A섹션(생활필수품)에 관한 특허문헌에 대하여 학습하여 SVM과 비교한 결과, 분류 속도는 전체적으로 향상되었으나, 분류 정확도는 SVM(84.97%)보다 낮은 83.66%를 나타내고 있다.

[73]의 연구에서는 *k*NN 알고리즘을 이용하여 복수의 융합기술에 대한 특허문헌에 대하여 복수의 IPC코드를 부여하는 방안을 제시하였으며, [74]의 연구에서는 웨어러블 디바이스와 사물인터넷의 융합기술과 관련하여 727건 특허문헌에 있어 509건을 학습데이터로 218건을 검증데이터로 SVM, 나이브 베이지안, *k*NN을 이용하여 90%이상의 분류 성능을 확인하였다.

[75]의 연구에서는 특허 자동분류에 있어서 비교적 적게 출현하는 용어의 경우, 일반적으로 무시되거나 노이즈로 처리되어 학습에 사용되지 않지만, 저빈도의 용어들을 특징으로 이용하여 분류의 효율을 높이고, 학습속도를 향상시킬 수 있도록 연구하였다. 또한, [76]의 연구에서는 특허문헌에 대하여 기술분야 중에서 전기공학, 기기, 화학, 기계공학 및 기타기술 5가지 기술분야로 분류로 정리하여 발명의 명칭에 사용된 단어의 출현빈도 변화를 계산하여 기술동향 분석을 연구하였다.

2.3 신경망이론

신경망 이론은 인간이 학습해가는 뇌신경세포 회로를 표본으로 데이터로부터 반복적인 학습과정을 거쳐 패턴을 찾아가는 이론이다[77]. 신경망은 뉴런과 노드로 구성되어 주로 감독학습에 이용되며, 입력변수와 출력변수 사이 또는 복수의 입력변수 사이에 존재하는 복잡한 관계를 파악할 수 있는 알고리즘이다.

신경망이론은 예측이나 분류를 목적으로 감춰진 패턴을 구하거나, 문서의 분류 또는 기존 데이터를 사용하여 노드들 간의 연결에 가중치를 조정하는 학습 등에 종종 사용된다[78]. 이러한 신경망의 종류에는 기본적으로 입력항목을 표시하는 입력층과 입력항목 간의 조합을 통해 나온 결과를 출력하는 출력층(output layer)으로 이루어진 단층신경망(single layer neural network)과 그림 2.1과 같이, 입력층과 출력층 그리고 입력층과 출력층 사이에 하나 이상의 은닉층 이루어진 다층신경망(multi-layer neural network)으로 구분된다.

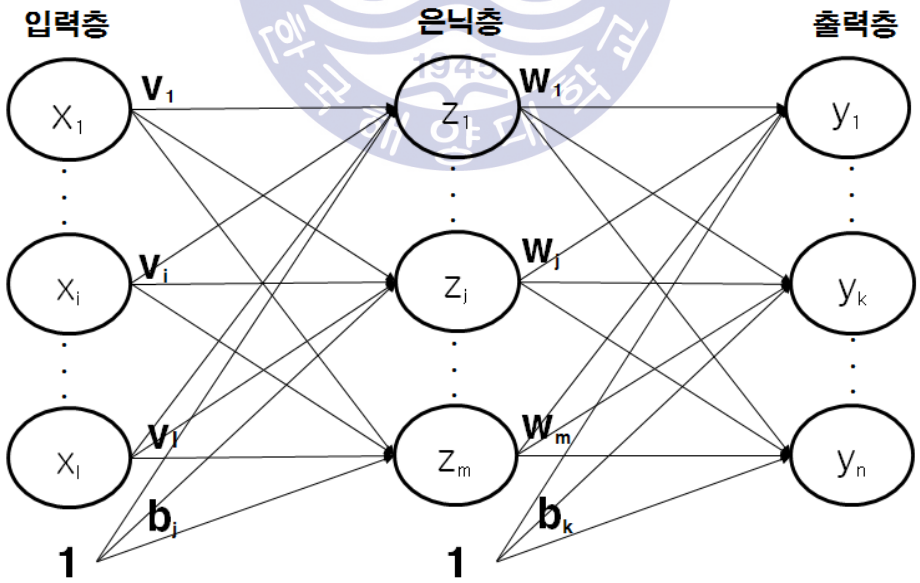


그림 2.1 인공신경망 구조

Fig. 2.1 A structure of an artificial neural network

각 층은 하나 이상의 노드를 포함하는데, 입력층에는 각 입력변수에 대응되는 수의 노드로 구성된다. 은닉층은 입력층으로부터 전달된 입력변수 값들을 이용한 선형결합을 비선형함수로 처리하여 출력층 또는 다른 은닉층에 전달한다. 출력층은 출력변수에 대응되는 뉴런으로서 분류모형에서는 분류되는 클래스의 수 만큼의 출력노드가 생성된다.

이렇게 구성된 신경망 구조에서 학습을 한다는 것은 입력변수값이 입력층에 입력된다. 그리고 목적으로 하는 목표값이 출력층을 통해 결과가 나오도록 각 계층의 노드들을 연결하는 가중치를 설정한다는 것을 의미한다.

외부에서 입력층에 입력변수값이 주어지면, x_i는 이를 나타내는 입력값이고, 외부의 바이어스(bias)인 임계값으로 1.0이 주어진다. v_i는 해당 노드에서의 입력값에 주어지는 가중치이며, b_j는 임계값에 대한 가중치이다. 이와 같이 결정된 값의 선형결합에 의한 결과가 은닉층인 z_j노드의 입력값으로 식 (2.5)로 표현된다. 여기서, 은닉층에서의 $f(\cdot)$는 softmax 함수나 ReLu 함수를 사용한다.

$$z_j = f(\sum_{i=1}^{l}(v_i \times x_i) + b_j) \tag{2.5}$$

그리고 출력층은 식 (2.6)으로 표현되며, z_j는 은닉층의 각 노드에서 출력층으로 입력되는 입력값이고, 외부의 임계값으로 1.0이 주어진다. 그리고 w_j는 은닉층의 각 노드에서 주어지는 가중치이며, b_k는 출력층에 입력되는 임계값에 대한 가중치이다.

$$y_k = \sigma(\sum_{j=1}^{m}(w_j \times z_j) + b_k) \tag{2.6}$$

여기서, $\sigma(\cdot)$는 각 노드에서의 활성화함수로 s자 모양의 함수를 주로 사용한다. 특히, 출력층에서의 $\sigma(\cdot)$는 softmax 함수나 ReLu 함수가 주로 사용된다.

이와 같은 형태의 신경망을 전방향(feedforward) 신경망이라고 하며, 이러한 전방향 신경망에서는 출력된 출력값이 학습의 목적에 부합하지 못하는 경우에 이를 반영하지 못하는 단점이 있다.

이를 극복하도록 가중치를 조정하는 알고리즘으로 가장 대표적인 신경망 알고리즘이 역전파(back-propagation) 신경망 알고리즘이다. 일반적으로 전방향 신경망을 훈련시키기 위한 알고리즘 중 훈련방법에 별도의 언급이 없으면 역전파 신경망 알고리즘을 의미한다[79].

역전파 신경망 알고리즘은 입력층에 값을 주면 은닉층에 전달되어 설정된 가중치를 기반으로 연산된 결과를 출력층에서 신호를 출력하게 되는데, 실제 출력층에서 출력된 출력값과 목표값사이에는 오차가 발생하게 되는데 이러한 오차를 감소시키도록 가중치를 조정하게 된다[80].

가중치들은 식 (2.7)을 이용한 확률적 기울기 하강법(stochastic gradient descent)을 통해 조정될 수 있다. 여기서, $\triangle w_i(t+1)$은 $t+1$번째의 가중치이고, $\triangle w_i(t)$는 t번째의 가중치이며, η는 학습률(learning rate)이고, E는 오차함수이다.

$$\triangle w_i(t+1) = \triangle w_i(t) + \eta \frac{\partial E}{\partial w_i} \tag{2.7}$$

오차함수의 선택은 학습의 형태와 활성화함수에 따라 결정되는데, 예를 들어, 다중 클래스 분류 문제에서 지도학습을 수행하는 경우에는 활성화함수와 오차함수로서 각각 softmax함수와 교차 엔트로피 함수(cross entropy function)가 결정된다. softmax 함수는 식 (2.8)과 같다. 여기서, p_j는 클래스 확률(class probability), 즉 어느 분류에 속할 확률을 나타낸다. 그리고 x_i와 x_k는 각각 유닛 j로의 전체 입력(total input)과 유닛 k로의 전체 입력을 나타낸다.

$$p_j = \frac{\exp(x_j)}{\sum_{k=1}^{n} \exp(x_k)} \tag{2.8}$$

교차 엔트로피는 식 (2.9)와 같다. 이때, $p_j^{'}$는 출력 유닛 j에 대한 목표 확률(target probability)을 나타내며, p_j는 해당 활성화함수를 적용한 이후의 j에 대한 확률 출력(probability output)이다[81].

$$E = -\sum_{j=1}^{n} p_j^{'} \log(p_j) \tag{2.9}$$

오차함수에 대하여 가중치로 편미분한 일차편미분벡터(gradient)가 0이 되면 오차는 최소화된다[82]. 이와 같은 역전파 알고리즘과 확률적 기울기 하강법을 이용하는 것은 국지적인 최적화를 이루기가 쉽고, 그 구현의 용이한 특징이 있어 다른 알고리즘에 비해 선호된다.

2.4 특허문헌의 특징

기술문헌 중에서 각 국가 특허청에서 발행하는 특허문헌은 다른 기술문헌들에 비해 그 내용이 비교적 객관적이며 체계적으로 서술된다[83]. 이는 국제화시대에 맞춰 특허문헌에 대하여도 일정한 형식으로 통일화된 결과이다. 특허문헌의 서지사항에는 출원일, 공개일, 등록일, 발명자, 출원인 및 특허심사를 위한 선행기술조사 문헌 및 특허분류코드 등을 제공하는데, 그 텍스트 형식은 표 2.1과 같이, 서로 다른 특징을 가지고 있다.

표 2.1 특허문헌의 텍스트 형식
Table 2.1 The text format of patent documents

숫자	출원번호, 출원일, 공개번호, 공개일, 등록번호, 등록일, 우선권주장번호 등
고유명사	출원인, 발명자, 대리인 등
일반 텍스트	발명의 명칭, 요약, 상세한 설명
특정스타일의 텍스트	청구항

따라서 이러한 특정 형식의 서지사항을 가진 특허문헌은 기술의 발전 흐름과 특정 기업 혹은 연구소 및 주요 연구원들의 연구개발 방향 등을 쉽게 분석할 수 있는 장점이 있다[84]. 실제 특허문헌에서는 신문기사 및 웹 페이지 등과 같은 일반문서를 대상으로 하는 기계학습을 이용한 유사도나 분류를 함에 있어 유사한 점도 많으나 특허문헌의 형식에 일부 차이점이 있어 이에 대한 고려가 필요하다.

일반 문서와 달리 특허 문헌에서는 해당 발명 기술의 특징을 보다 더 정확하게 표현하기 위하여 이미 일반화된 단어들을 결합시켜 새로운 단어를 창조해 내는 경우도 있다[85]. 예를 들어, 한국특허 10-1490518의 발명의 명칭은 '광역급행버스의 좌석예약방법 및 상기 방법에 의한 광역급행버스의 좌석예약장치'에서 사용된 '광역급행버스'는 일반적으로 경기도와 서울 간 운행되는 버스에 사용되는 단어이나, 본 발명에서는 준비된

좌석의 수만큼 승객이 탑승할 수 있는 버스를 모두 포함하는 개념으로 새로운 의미를 부여하였다[86].

그리고 특허문헌의 명세서의 내용에서 일반적인 용어를 사용하게 되면, 특허권자의 권리를 판단하는데 있어 가끔 권리범위가 한정되는 경우가 발생하므로, 기술용어는 그 의미가 상당히 넓고 포괄적인 용어를 사용하도록 권장한다. 예를 들어, 모바일 단말기에 관한 발명에서 사용자가 단말기의 터치패드를 통해 데이터를 입력받아 전송하고, 입력받은 데이터를 단말기의 중앙처리장치를 거쳐 단말기의 화면에 표시하는 발명에 대하여 발명자가 특허문헌에 자신의 권리를 주장하고자 하는 청구항에 기재할 때는 각각의 장치에 대한 포괄적인 용어인 입력부, 전송부, 처리부 및 표시부와 같이 최대한 그 의미를 확장시키는 용어를 사용한다.

또한, 일반문서에서 사용되는 단어의 경우에는 누구나 쉽게 이해할 수 있는 단어 범용의 단어가 사용되고 있지만, 특허문헌의 경우에는 특정 기술분야에서만 사용되는 전문용어가 많다는 점도 지적되고 있다[87]. 다시 말해 특허 문헌은 일반적인 용어가 아닌 특정 기술분야에 한정되어 사용되는 용어를 사용하기도 하고, 권리범위를 보다 넓게 해석할 수 있도록 포괄적인 의미의 용어를 사용하기도 하며, 일반적인 용어를 결합을 통해 새로운 용어를 창조하기도 한다.

특허문헌은 특허문헌 양식에 맞게 문서의 형식이 정형화 된 다양한 서지정보를 포함하고 있다. 또한 특허문헌은 일반문서와 달리 분류 코드에 따라 매우 계층적이고 복잡하게 분류가 가능한 특징이 있다.

특허문헌은 기술의 권리를 정확히 표현하기 위해 일반 문서에 비해 표현이 비교적 길게 서술되어 있다. 그리고 기술이 속하는 분야에 따라 문서의 길이가 다양하게 존재한다. 또한 대부분의 특허문헌은 기술의 특징을 나타내는 도면을 포함하고 있다.

특허문헌은 발명에 대하여 독점적이고 배타적 권리를 받기 위해 타인의 권리와의 관계를 명확히 하고자 특허문헌과 관련된 날짜가 매우 명확하고, 제목, 요약, 특허청구항 및 설명이 의미적으로 연관성이 매우 높다. 특허의 핵심적인 내용은 발명의 명칭 및 청구항에 나타나며, 해당 특허와 관련성이 높은 선행기술문헌에 대한 정보가 포함되는 경우가 많고, 특허의 핵심적인 내용을 대표하는 키워드가 기재되지 않는다는 특징이 있다 [88].

제 3 장 핵심 특허문헌 추출 방법

3.1 핵심 특허문헌 추출시스템의 개요

본 논문에서 제안하는 핵심 특허문헌 추출시스템은 그림 3.1에서와 같이, 특허동향분석의 절차 중 유효 특허문헌의 추출, 세부기술별 분류 및 핵심 특허문헌의 추출 절차를 정보검색 및 신경망 알고리즘을 이용하여 자동 수행이 가능하도록 한다. 먼저, 핵심 특허문헌을 추출하기 위하여 기술요약서에 기재되어 있는 색인어를 기반으로 검색식을 작성하여 특허검색서비스를 통해 후보 특허문헌 목록을 구한다.

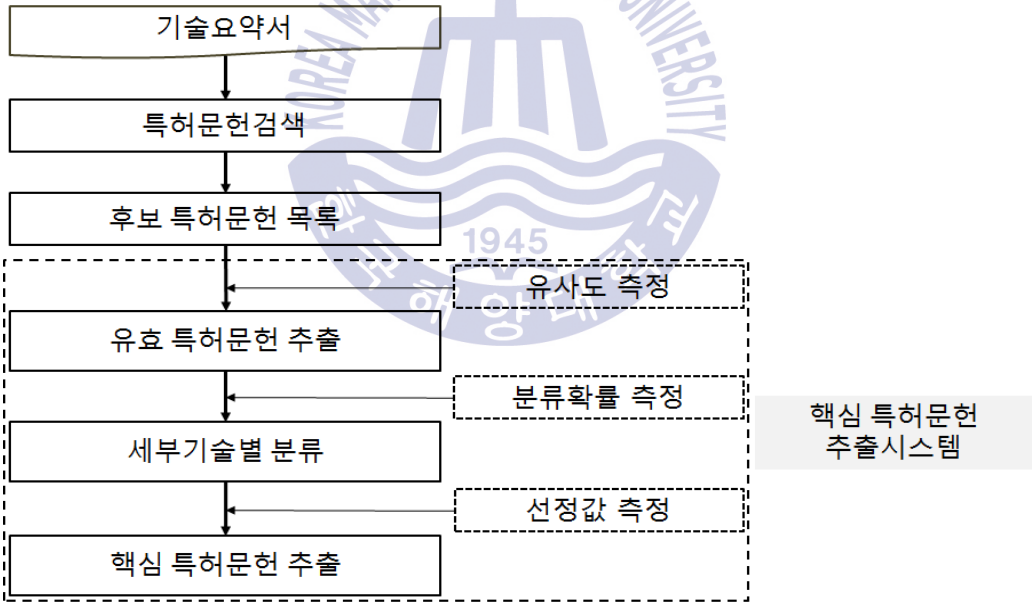


그림 3.1 핵심 특허문헌 추출시스템을 이용한 특허동향분석

Fig. 3.1 The patent trend analysis using core patent document extraction system

핵심 특허문헌 추출시스템은 후보 특허문헌 목록을 대상으로 문서간의 유사도를 측정하여 유효 특허문헌을 추출하고, 추출된 유효 특허문헌의 기술별 분류확률을 측정하여 상대적으로 높은 분류확률로 유효 특허문헌을 분류하며, 분류된 유효 특허문헌의 항목들에 기재된 값과 우선순위가중치를 연산한 선정값을 순위화시켜 핵심 특허문헌을 추출한다. 이와 같이 핵심 특허문헌 추출시스템을 통하여 얻어진 전체 유효 특허문헌과 기술분류별 유효 특허문헌 및 핵심 특허문헌을 이용하여 전체 연도별 특허출원동향, 주요 시장국 내외국인 특허출원현황, OS Matrix 분석, IP 장벽도 분석 등의 특허동향분석 보고서를 작성하게 된다.

핵심 특허문헌 추출시스템은 유효 특허문헌의 추출 단계, 세부기술별 분류 단계, 핵심 특허문헌 추출 단계가 순차적으로 이루어지며, 그림 3.2는 각 단계에서 이루어지는 특허문헌의 순위화 절차를 나타낸다.

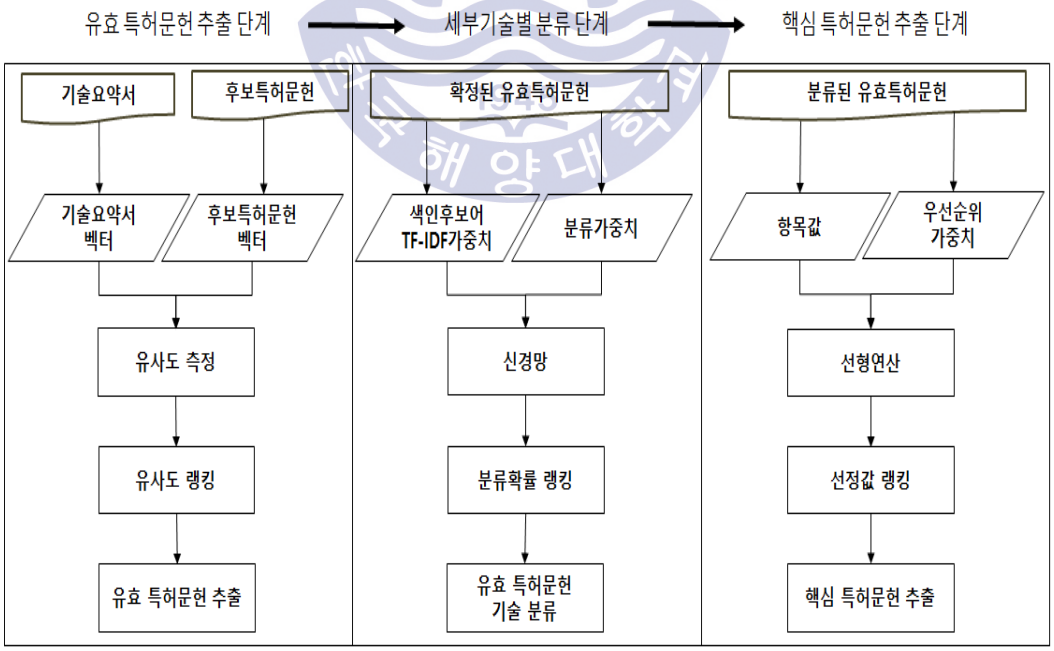


그림 3.2 추출시스템의 특허문헌 순위화 절차

Fig. 3.2 The ranking procedure of our proposed system

핵심 특허문헌 추출시스템의 유효 특허문헌 추출 단계는 기술에 대한 정보인 기술요약서의 문서벡터와 후보 특허문헌의 문서벡터를 각각 구한다. 그리고 이들 문서벡터들을 입력데이터로 문서간의 유사도를 측정하여 유사도가 높은 순으로 순위화하여 유효 특허문헌을 추출한다.

세부기술별 분류 단계는 유효 특허문헌으로 추출된 문서를 대상으로 각 유효 특허문헌에 기재된 색인후보어에 대한 TF-IDF가중치와 분류와 관련된 색인후보어에 대한 분류가중치를 구한다. TF-IDF가중치와 분류가중치를 신경망 알고리즘의 입력데이터로 이용하여 세부기술별 각 유효 특허문헌의 분류확률을 순위화하고, 상대적으로 분류확률이 높은 세부기술을 선택함으로써 유효 특허문헌을 기술별로 분류한다.

핵심 특허문헌 추출 단계는 분류된 유효 특허문헌에 기재되어 있는 항목들 중에서 특허조사원이 특허문헌의 중요도를 판단하는데 사용하는 항목인 등록여부, 패밀리국가수 등의 항목들과 유사도 항목에 대하여 우선순위를 부여한다. 그리고 특허조사원이 선정한 항목들의 값과 우선순위에 따른 우선순위가중치를 입력으로 이들을 선형 연산하여 구한 선정값을 순위화하여 핵심 특허문헌을 추출한다.

3.2 유효 특허문헌 추출

3.2.1 자질추출

문서와 문서간의 유사도를 측정하는 방법에는 하나의 문서가 다른 문서와 유사하다고 판단할 수 있는 특징이 있어야 한다. 그리고 이러한 특징들이 높은 유사성을 나타낼 때 유사하다고 할 수 있다.

앞서 기술요약서로부터 추출한 색인어와 이들의 동의어를 기반으로 작성한 검색식을 통해 후보 특허문헌 목록을 획득하였다. 불리언 검색의 특징상 검색식에 포함된 색인어와 동의어는 해당 특허문헌에 반드시 포함되어 있다. 이에 검색식을 구성하는 색인어는 연구개발기술에 대한 기술요약서와 검색된 특허문헌간의 유사도를 판단하는 특징으로서 한계가 있다.

이에 본 논문에서는 연구개발 기술의 내용과 보다 유사한 특허문헌을 추출하기 위하여 기술요약서 자체와 각 특허문헌간의 유사도를 판단하고자 한다. 문서의 자질로서 문서의 내용을 잘 표현하고 있는 문서벡터, 즉 기술요약서의 문서 벡터와 각 특허문헌의 문서벡터를 비교하여 유사도를 측정한다.

3.2.2 기술요약서의 자질추출

기술요약서의 문서벡터는 색인후보어를 기준으로 각 색인후보어에 대한 가중치를 산출한다. 색인후보어에 대한 가중치 산출 방법으로 일반적으로 가장 널리 알려진 TF-IDF 가중치를 적용한다. 기술요약서의 각 색인후보어에 대한 TF값은 식 (3.1)과 같다.

$$tf_{i,j}^{t} = \frac{f_{i,j}^{t}}{\sum_{i=1}^{n} f_{i,j}^{t}} \tag{3.1}$$

여기서, $tf_{i,j}^t$는 기술요약서 d_j^t에 대한 특정 색인후보어 t_i의 출현율이며, $f_{i,j}^t$는 기술요약서 d_j^t에서 색인후보어 t_i의 출현 빈도수이고, $\sum_{i=1}^{n} f_{i,j}^t$는 기술요약서 d_j^t에 기재된 모든 색인후보어들의 출현 빈도수이다. 이는 각 색인후보어의 정형화(formalization)를 위해 기술요약서내 하나의 색인후보어의 출현 빈도수를 모든 색인후보어의 출현빈도수로 나누어 구한다. 기술요약서에서 색인후보어에 대한 IDF값은 식 (3.2)와 같다.

$$idf_i^t = \log \frac{|D^t|}{|\{d_j^t \mid t_i \in d_j^t\}|} \tag{3.2}$$

여기서, idf_i^t는 전체 기술요약서 d_j^t에 대한 색인후보어 t_i의 기여율이다. $|D^t|$는 기술요약서의 총 개수이고, $|\{d_j^t \mid t_i \in d_j^t\}|$는 색인후보어 t_i가 포함된 기술요약서 d_j^t의 개수이다. 그리고 각 색인후보어에 대한 TF-IDF가중치는 식 (3.3)과 동일하다.

$$w_{i,j}^t = tfidf_{i,j}^t = tf_{i,j}^t \times idf_i^t \tag{3.3}$$

다만 기술요약서의 문서가 하나이기 때문에 정형화를 위해 별도로 로그를 취할 필요가 없다. 따라서 IDF값은 1이다. 따라서 기술요약서의 문서벡터는 각 색인후보어의 TF-IDF가중치는 TF값과 같다. 표 3.1은 기술요약서의 각 색인후보어의 TF-IDF가중치 예시이다.

표 3.1 기술요약서내 색인후보어들의 TF-IDF 예시

Table 3.1 An example of TF-IDF for index candidates in technical summary

색인후보어	TF-IDF(TF값)
단어1	0.050420
단어2	0.042017
⋮	⋮
단어n	0.004202

이와 같이 구해진 기술요약서의 각 색인후보어에 대한 TF-IDF가중치는 기술요약서의 자질인 문서벡터(d_j^t)에 해당한다. 문서벡터(d_j^t)는 식 (3.4)로 표현된다.

$$d_j^t = (w_{1,j}^t, w_{2,j}^t, \cdots, w_{n,j}^t) = (0.050420, 0.042017, \cdots, 0.004202) \quad (3.4)$$

3.2.3 특허문헌의 자질추출

특허검색서비스에서 검색식을 통해 확보한 후보 특허문헌 목록은 텍스트의 csv 또는 excel 형태의 파일이다. 후보 특허문헌 목록의 각 열에는 검색된 하나의 특허문헌에 대한 출원번호, 공개번호, 등록번호 등의 번호정보, 출원인, 발명자 등의 인적정보, 출원일, 등록일, 공개일, 우선권주장일 등의 날짜정보, 발명의 명칭, 요약, 대표청구항, 청구항 등의 기술내용정보, IPC코드 분류, FI분류, F-TERM 등의 분류정보, INPODOC 패밀리수, 평가점수, 피인용수 등의 기타정보 등 많은 정보를 나타내고 있다.

후보 특허문헌 목록에는 특허문헌의 특징을 나타내는 다수의 항목들이 포함되어 있다. 하지만 특허문헌의 자질을 추출하기 위하여 텍스트 기반의 항목만이 필요하고, 또 시스템의 부하를 줄이기 위해 검색된 후보 특허문헌 목록에 대한 수정이 필요하다. 그림 3.3과 같이, 후보 특허문헌 목록의 각 특허문헌은 기술내용에 해당하는 항목인 발명의 명칭, 요약, 대표청구항, 청구항과 키 값으로 출원번호를 제외한 항목을 삭제하였다.

번호	발명의 명칭	요약	대표 청구항	전체 청구항
KR2009003	항체의 F c 영역에 특이적 결	본 발명은 항체의 Fc 영역에	1 항체의 Fc 영역 결합	1 항체의 Fc 영역 결합 펩타이드
KR2000003	인체의 피질 흉선 세포 및 백혈	본 발명은 인체의 피질 흉선	1 서열 1에 기재된 아미	1 서열 1에 기재된 아미노산 서
KR1995070	비오틴/아비딘-금속 단백질 콘	본 발명은 환자의 병변을 검	1 환자에게, (a)표적 조	1 환자에게, (a)표적 조성물을
KR1992070	감염성 병변 및 염증성 병변을	본 발명의 백혈구 유착 병소	1 백혈구 유착 병소를	1 백혈구 유착 병소를 타게팅하
KR2011700	음이온성 인지질 및 아미노인지	종양 혈관구조 및 바이러스	110% 혈청의 존재하에	110% 혈청의 존재하에 수행된
KR1986000	메탈로티오네인과 생물학적 활	<목적>진단 및 치료를 목적	1 생물학적 활성 분자(	1 생물학적 활성 분자(BAM)를
KR2009700	F Z D10에 대한 종양-표적화	본 발명은 프리즐드 상동체	1 서열 15, 17 및 19에	1 서열 15, 17 및 19에 제시된
KR2011001	개선된 플루오르-18 표지를 위	본 발명은 하기 화학식 1로	1하기 화학식 1로 표시	1하기 화학식 1로 표시되는 F-
KR2004701	피브로넥틴의 E D-B 도메인	본 발명은 방사성핵종을 결	1 aa) 이하의 표에 보여	1 aa) 이하의 표에 보여진 상보
KR2016700	급성 골수성 백혈병의 진단용	폴리펩티드 특히 항체 또는	1AML 아형 M2 특이적	1AML 아형 M2 특이적 진단용
KR2015703	급성 골수성 백혈병 세포를 표	AML 아세포의 표면에 결합	1아미노산 서열을 포함	1아미노산 서열을 포함하는 항
KR2015702	HGF에 대한 항체 및 이를 포함	본 발명은 HGF에 대한 결합	1하기를 특징으로 하는	1하기를 특징으로 하는 항-인간
KR2015703	전립선암의 바이오마커 검출에	본 발명은 (a) 개체로부터	1개체에서 전립선암-관	1개체에서 전립선암-관련 질환
KR2015701	압타머 및 이의 용도{APTAME	관심 생물지표, 이를 테면	1미소낭포에 결합하고	1미소낭포에 결합하고, a) 표 1

그림 3.3 후보 특허문헌 목록의 기술내용 항목 예시

Fig. 3.3 An example of technical fields in candidate patent document list

본 논문에서는 이러한 특허문헌의 자질에 부합하는 TF-IDF가중치, 항목별가중치 및 비색인어가중치를 제시하고, 이들 가중치를 반영하여 유사도를 측정한다. TF-IDF가중치는 후보 특허문헌 목록에서 각 특허문헌의 기술내용에 기재되어 있는 항목인 발명의 명칭, 요약, 대표청구항, 청구항의 필드로부터 기술요약서에서 추출한 색인후보어의 TF-IDF가중치이다. TF-IDF가중치를 구하기 위한 TF값은 식 (3.5)와 같다.

$$tf_{i,j}^{p} = \frac{f_{i,j}^{p}}{\sum_{i=1}^{n} f_{i,j}^{p}} \tag{3.5}$$

여기서, $tf_{i,j}^{p}$는 특허문헌 d_j^p에 대한 특정 단어 t_i의 출현율이다. $f_{i,j}^{p}$는 특허문헌 d_j^p에서 색인후보어 t_i의 출현 빈도수이며, $\sum_{i=1}^{n} f_{i,j}^{p}$는 특허문헌 d_j^p에서 모든 색인후보어들의 출현 빈도수이다. 그런데 만약 색인후보어1의 경우, 첫 번째 특허문헌의 요약에 1회 기재되고 특허문헌에 기재된 색인후보어의 수가 전체 10개인 경우, TF값은 0.1이다.

그러나 다음 특허문헌에서 같은 색인후보어가 역시 1회 요약에 기재되고 해당 특허문헌에 출현된 색인후보어의 수가 20인 경우에는 TF값은 0.05가 된다. 이는 유사도를 측정하는데 있어 출현된 색인후보어의 빈도수가 높음에도 불구하고, 유사도가 적게 나타나는 문제, 즉 TF값의 정형화에 문제가 있다. 이러한 문제를 해결하기 위해서 모든 색인후보어들의 출현 빈도수는 색인후보어의 수로 정한다.

다음으로 특허문헌의 IDF값은 식 (3.6)이다. 여기서, idf_i^p는 전체 특허문헌 d_j^p에 대한 단어 t_i의 기여율이다. $|D^p|$는 특허문헌의 총 개수로 후보 특허문헌 목록의 총수인 열의 수이고, $|\{d_j^p \mid t_i \in d_j^p\}|$는 색인후보어 t_i가 포함된 특허문헌 d_j^p의 개수이다. 그리고 색인후보어에 대한 각 특허문헌의 TF-IDF가중치는 식 (3.7)과 같다.

$$idf_i^p = \log \frac{|D^p|}{|\{d_j^p \mid t_i \in d_j^p\}|} \tag{3.6}$$

$$w_{i,j}^{p} = tfidf_{i,j}^{p} = tf_{i,j}^{p} \times idf_i^p \tag{3.7}$$

그런데 하나의 특허문헌은 텍스트 필드 중 발명내용과 관련하여 4개의 항목을 포함하고 있어서, 각 항목 중요도에 따라 다른 가중치를 주는 것이 필요하다. 그래서 두 번째 가중치로서 항목별가중치를 제안한다. 항목별가중치는 특허기술의 내용과 관련된 4개의 텍스트 항목인 발명의 명칭(title), 요약(abstract), 대표청구항(first claim), 청구항(claim)에 기재된 색인후보어에 대하여 각 항목이 가지는 특징에 따라 서로 다른 가중치를 부여하는 것이다.

발명의 명칭은 해당 특허기술을 가장 잘 표현하는 내용을 간단하고 명확하게 기재하는 것으로, 발명의 내용을 가장 함축적으로 표현하고 있어서 기술내용의 특징을 잘 나타내고 있으므로 가중치가 가장 높다. 요약은 특허기술의 구체적인 내용을 요약하여 기술한 것으로, 특허기술의 내용을 개괄적으로 표현하고 있어서 기술분야나 내용의 특징을 확인할 수 있으므로 가중치가 발명의 명칭 다음으로 높다.

대표 청구항은 특허기술의 권리범위를 표현하는 청구항 중에서 독립 청구항으로 가장 기본적인 권리의 범위를 간단하고 명확하게 기재한다. 그러나 내용이 한정적으로 표현되어 전체 특허기술을 표현하는데 한계가 있으므로 가중치가 낮다. 청구항은 대표청구항에 기반한 부가적인 기술내용을 간단하게 명확하게 기재하는 것으로 가중치가 가장 낮다.

이와 같이 항목별가중치는 하나의 특허문헌에 포함된 특허기술의 내용과 관련된 발명의 명칭, 요약, 대표청구항, 청구항의 순으로 부여하고자 한다. 그리고 각 항목에 따른 항목별가중치는 표 3.2와 같이 여섯 세트를 제시하여 각 세트별 항목별가중치를 이용하여 유사도를 측정하여 추출된 유효 특허문헌을 비교하고자 한다.

표 3.2 항목별가중치 세트

Table 3.2 The sets of field weights

항목별가중치 세트	발명의 명칭 (α)	요약 (β)	대표청구항 (γ)	청구항 (δ)	항목별 가중치간격
첫 번째 세트	1.00	0.95	0.90	0.85	0.05
두 번째 세트	1.00	0.90	0.80	0.70	0.10
세 번째 세트	1.00	0.85	0.70	0.55	0.15
네 번째 세트	1.00	0.80	0.60	0.40	0.20
다섯 번째 세트	1.00	0.75	0.50	0.25	0.25
여섯 번째 세트	1.00	0.70	0.40	0.10	0.30

실험 대상은 검색된 후보 특허문헌 목록을 대상으로 유효특허의 수와 정확률이 높은 항목별가중치를 적용하여 특허문헌의 자질에 반영한다. 그리고 항목별가중치 세트는 0에서 1 사이의 범위에서 정하였다. 왜냐하면 TF-IDF가중치를 구하는데 있어 로그를 이용하여 0과 1사이의 값으로 한정하였기 때문이다.

그리고 유사여부를 판단하는 기술 내용의 중요도는 발명의 명칭에서 청구항의 순이지만, 색인후보어가 발견될 확률은 비교적 기재된 단어가 많은 청구항부터 발명의 명칭 순이기 때문에 각 세트들간의 항목별로 항목별가중치 값의 간격이 큰 순서는 표 3.3과 같이, 청구항, 대표청구항, 요약, 발명의 명칭순으로 정하여 진다.

표 3.3 항목별가중치 세트간 간격

Table 3.3 The intervals among field weight sets

항목별가중치	발명의 명칭 (α)	요약 (β)	대표청구항 (ϒ)	청구항 (δ)
세트간 가중치 간격	0.00	0.05	0.10	0.15

따라서 선정된 항목별가중치를 적용하면, 식 (3.5)에서 TF값이 각 항목별 가중치가 적용된 식 (3.8)로 변경된다. 여기서 $ttf_{i,j}^{p}$는 하나의 특허문헌에서 색인후보어가 기재된 각 항목별 TF값의 합이다. 그리고 각 항목별 TF값은 그 항목에 따라서 발명의 명칭은 $f_{i,j}^{t}$, 요약은 $f_{i,ji}^{a}$, 대표청구항은 $f_{i,j}^{f}$, 청구항은 $f_{i,j}^{c}$로 표기하였으며, 가중치는 발명의 명칭, 요약, 대표청구항, 청구항이 각각 α, β, γ, δ로 나타낸다.

$$ttf_{i,j}^{p} = \frac{(f_{i,j}^{t} \times \alpha + f_{i,j}^{a} \times \beta + f_{i,j}^{f} \times \gamma + f_{i,j}^{c} \times \delta)}{\sum_{i=1}^{n} f_{i,j}^{p}} \tag{3.8}$$

그리고 IDF값은 문서를 기준으로 하는 값으로 항목별 차이는 의미가 없다. 따라서 TF-IDF가중치는 식 (3.7)에서 식 (3.9)로 변경된다.

$$w_{i,j}^{p} = tfidf_{i,j}^{p} = ttf_{i,j}^{p} \times idf_{i}^{p} \tag{3.9}$$

세 번째 가중치는 색인후보어 중에서 기술요약서에 기재되었지만, 특허검색의 검색식에 포함되지 않았던 색인후보어에 대하여 가중치를 부여한다. 이는 현재 유효 특허문헌이 특허검색의 특징인 불리언 검색을 통해 검색되었기 때문에 그 연구개발 기술과 관련성이 낮은 특허문헌도 포함될 수 있는 한계를 극복한다.

다시 말해서 연구개발 기술을 표현하는 색인후보어 중 일부만 검색식에 포함되었기 때문에, 색인어 외의 색인후보어를 이용하게 되면 기술요약서의 기술내용과 훨씬 더 유사한 특허문헌을 찾을 확률이 높아지게 된다. 따라서 검색식에 포함되지 않았던 색인후보어를 비색인어라고 하고, 이 가중치를 비색인어가중치(w_i^n)로 정의한다.

따라서 최종적인 특허문헌의 문서벡터의 가중치는 식 (3.10)과 같이 표현된다. 여기서 비색인어가중치(w_i^n)는 색인후보어 i가 색인어인 경우에는 1.0이고, 색인어가 아닌 경우에는 1.5이다.

$$w_{i,j}^p = w_{i,j}^p \times w_i^n \tag{3.10}$$

표 3.4는 특허문헌 색인후보어에 대한 TF값, 항목별가중치, IDF값, TF-IDF가중치, 비색인어가중치를 통해 구해지는 색인후보어들의 가중치 예시이다. 예를 들어 색인후보어1의 경우, 첫 번째 특허문헌의 요약에만 1회 기재되고 전체 색인후보어의 수가 10개인 경우, TF값은 0.1이고, 항목별가중치는 0.8로서, TTF값은 0.08이며, IDF값은 특허문헌 1000건 중의 2건에만 기재되어 log(1000/2) = 2.69897로 TF-IDF가중치는 0.2159176이고, 비색인어에 해당하므로 비색인어 1.5가 곱해지면 문서벡터가 0.3238764가 된다.

표 3.4 색인후보어별 가중치 예시

Table 3.4 An example of weights by index candidates

색인후보어	$tf_{i,j}^{p}$	$ttf_{i,j}^{p}$	idf_{i}^{p}	$tfidf_{i,j}^{p}$	w_{i}^{n}	$w_{i,j}^{p}$
색인후보어1	0.1	0.08	2.69897	0.215918	1.5	0.323876
색인후보어2	0.2	0.36	2.15490	0.775764	1.0	0.775764
⋮	⋮	⋮	⋮	⋮	⋮	⋮
색인후보어10	0.1	0.04	2.39794	0.095918	1.0	0.095918

추가적으로 색인후보어2는 항목 중 명칭과 요약에 기재되고, 7건의 특허문헌에 기재된 경우의 예시이다. 그리고 마지막 색인후보어10의 경우는 청구항에 기재되고, 4건의 특허문헌에 기재된 경우의 예시이다. 표 3.4에 표시된 각 색인후보어 단어에 대하여 계산된 가중치의 조합이 각 특허문헌의 벡터로 식 (3.11)과 같다.

$$d_{j}^{p} = (w_{1,j}^{p}, w_{2,j}^{p}, \cdots, w_{n,j}^{p}) = (0.323876, 0.775764, \cdots, 0.095918) \tag{3.11}$$

3.2.4 유사도 측정

유사도를 측정하기 위하여 가장 일반적인 방법인 코사인 유사도를 적용한다. 코사인 유사도는 내적공간의 두 벡터간 각도의 코사인값을 이용하여 측정한 벡터간의 유사한 정도를 의미한다.

따라서 3.5.2절에서 계산된 기술요약서의 문서벡터(d_j^t)와 3.5.3절에서 계산된 특허문헌의 문서벡터(d_j^p)간의 스칼라곱를 이용한 코사인 유사도로 측정한다. 코사인 유사도의 측정방법은 식 (3.12)와 같다.

$$sim(d_j^t, d_j^p) = \frac{d_j^t \cdot \quad d_j^p}{|d_j^t| \cdot \quad |d_j^p|} = \frac{\sum_{i=1}^{n}(w_i^t \times w_i^p)}{\sqrt{\sum_{i=1}^{n}(w_i^t)^2} \times \sqrt{\sum_{i=1}^{n}(w_i^p)^2}} \tag{3.12}$$

여기서, d_j^t는 기술요약서의 각 색인후보어의 가중치 값의 조합이고, d_j^p는 특허문헌의 각 색인후보어의 가중치 값의 조합이다. 그리고 분자는 식 (3.13)과 같이, 기술요약서의 문서벡터(d_j^t)와 특허문헌의 문서벡터(d_j^p)간의 벡터내적(dot product)이다. 여기서 n은 색인후보어의 총 개수이다.

$$\sum_{i=1}^{n}(w_{i,j}^t \times w_{i,j}^p) = w_{1,j}^t \times w_{1,j}^p + w_{2,j}^t \times w_{2,j}^p + \cdots + w_{n,j}^t \times w_{n,j}^p \tag{3.13}$$

분모는 기술요약서의 문서벡터(d_j^t)의 크기와 특허문헌의 문서벡터(d_j^p)의 크기를 구해 곱하는 것으로 식 (3.14)와 같다. 이와 같이 계산된 기술요약서와 각 특허문헌간의 유사도 측정 결과를 바탕으로 특허문헌을 순위화함으로써, 기술요약서의 기술내용과 관련성이 낮은 특허문헌을 제거하여 특허동향분석에 필요한 유효 특허문헌을 획득할 수 있다.

$$\sqrt{\sum_{i=1}^{n}(w_{i,j}^t)^2} \times \sqrt{\sum_{i=1}^{n}(w_{i,j}^p)^2} = \sqrt{(w_{1,j}^t)^2 + (w_{2,j}^t)^2 + \cdots} \times \sqrt{(w_{1,j}^p)^2 + (w_{2,j}^p)^2 + \cdots} \times \cdots \tag{3.14}$$

3.3 세부기술별 특허문헌 분류

특허동향분석은 연구개발 기술의 대상과 조사 · 분석 기간 및 범위에 따라서 크게 A 타입의 특허전략보고서, B 타입의 특허분석보고서, C 타입의 특허동향보고서로 구분된다[89, 90]. 그리고 분석 및 조사를 수행함에 있어 보고서의 타입에 따라 B 타입의 경우에는 해당 기술을 다시 하위 기술 레벨의 중분류로, A 타입의 경우는 중분류를 다시 소분류까지 나누어 하위 기술분야에서의 분석 및 전략의 수립이 필요하다. 그러나 경우에 따라서는 타입에 관계없이 소분류까지 분류가 필요하다.

신경망 알고리즘은 기본적으로 입력변수값을 수신하는 입력층과 입력층에서 수신된 입력변수와 가중치의 선형결합을 입력받는 은닉층 및 은닉층의 출력을 받아 결과를 출력하는 출력층으로 세 계층 신경망구조를 가지고 있다. 역전파 알고리즘은 신경망 알고리즘의 대표적인 알고리즘으로, 입력값에 대하여 발생할 목표값과 실제 출력층에서 출력되는 출력값의 오차를 최소화하도록 은닉층의 가중치를 개선하는 알고리즘이다. 최근 많이 연구되고 있는 딥러닝은 이러한 역전파 신경망 알고리즘을 확장한 개념이다.

본 논문에서는 그림 3.4에서와 같이, 기존의 세 계층 구조의 신경망 구조에 2개의 은닉층을 추가하였다. 따라서 제안된 신경망구조는 입력층, 세 개의 은닉층, 출력층으로 다섯 계층으로 구성된다.

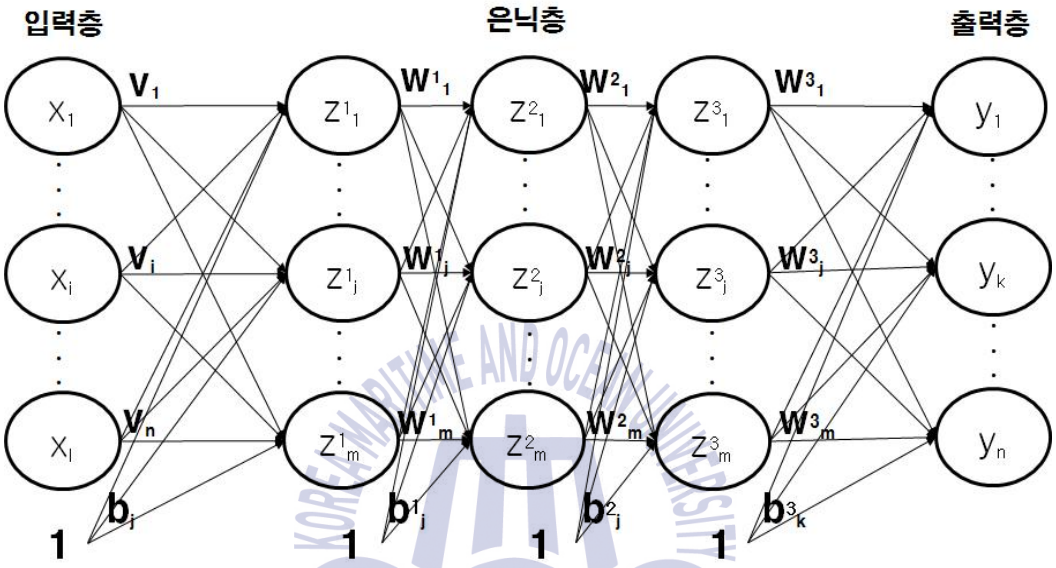


그림 **3.4** 제안된 인공신경망 구조

Fig. 3.4 The proposed structure of artificial neural networks

여기서 식 (3.15)와 같이, 입력층의 입력변수값 x_1 ~ x_l은 기술요약서에서 추출한 l개의 색인후보어들 각각이 하나의 특허문헌내에 기재되어 있는 단어와 일치하는 경우에 그 색인후보어들의 TF-IDF가중치이다.

$$x_i = tfidf_{i,j} = tf_{i,j} \times idf_i \tag{3.15}$$

예를 들어, 기술요약서의 색인후보어가 10개인 경우, 첫 번째 색인후보어를 특허문헌내에 기재된 단어와 비교하여 해당 색인후보어가 일치하는 단어가 있는 경우, 그 색인후보어의 TF값과 IDF값을 구한 후, TF값과 IDF 값을 곱한 TF-IDF가중치가 x_1의 입력변수 값이 된다. 그리고 두 번째 색인후보어에 대한 동일 특허문헌에서의 TF-IDF가중치

는 x_2의 입력변수 값이 된다. 이렇게 색인후보어 10개 단어의 TF-IDF가중치가 입력층 각각의 입력변수 값이 된다.

입력층에서 출력되어 제1은닉층에 입력되는 값인 z_j^1은 식 (3.16)과 같다. 여기서 분류가중치 v_i는 색인후보어 중에서 기술분류와 관련된 단어인 경우에는 기술분류를 위한 기계학습에 있어 특정 기술분류와 관련되어 있는 중요한 색인후보어에 해당한다. 따라서 기술분류와 관련된 색인후보어의 경우에는 v_i값으로 2.0를 할당하고, 그 외의 색인후보어의 v_i값은 1.0이다.

$$z_j^1 = f(\sum_{i=1}^{l}(v_i \times x_i) + b_j) \tag{3.16}$$

그리고 입력층에서의 바이어스(bias)인 임계값은 1.0이고, 임계값에 대한 가중치는 b_j이다. 그리고 은닉층에 해당하는 세 개 계층의 가중치(w_j^1 ~ w_j^3)는 초기값이 0이다. 제1은닉층에서 출력되어 제2은닉층에 입력되는 값인 z_j^2은 식 (3.17)과 같다. 제2은닉층에서 출력되어 제3은닉층에 입력되는 값인 z_j^3은 식 (3.18)과 같다.

$$z_j^2 = f(\sum_{i=1}^{l}(w_j^1 \times z_j^1) + b_j^1) \tag{3.17}$$

$$z_j^3 = f(\sum_{i=1}^{l}(w_j^2 \times z_j^2) + b_j^2) \tag{3.18}$$

제1은닉층, 제2은닉층 및 제3은닉층의 바이어스(bias)인 임계값은 역시 1.0로 주어진다. 여기서, 은닉층에서의 $f(\cdot)$는 ReLu 함수를 사용한다.

그리고 출력층은 식 (3.19)로 표기된다. 여기서, 출력층에서의 $\sigma(\cdot)$는 softmax 함수를 사용한다.

$$y_k = \sigma(\sum_{i=1}^{n}(w_j^3 \times z_j^3) + b_k^3) \tag{3.19}$$

출력층인 y_1 ~ y_n은 신경망 알고리즘을 통해 얻고자 하는 유효 특허문헌들의 기술분류에 해당한다. 예를 들어, 기술분류가 방사성 의약품, 영상장치, 유도약물인 경우에 출력층은 y_1 ~ y_3로 y_1은 방사성 의약품 관련 기술에 해당하고, y_2는 영상장치 관련 기술이며, y_3은 유도약물 관련 기술에 해당한다.

신경망 알고리즘에서 학습을 한다는 것은 입력층에 패턴을 입력하였을 때 목표로 하는 출력값이 나오도록 가중치를 설정하는 과정이다[91]. 따라서 은닉층의 가중치는 지도학습에 이용되는 훈련모집단에 속하는 유효 특허문헌을 통해 각각의 가중치값이 변경된다.

즉, 훈련모집단에 속하는 유효 특허문헌의 색인후보어에 대한 각각의 TF-IDF가중치인 입력변수 값과 신경망 알고리즘에 의해 출력된 기술분류에 해당하는 출력값 그리고 해당 유효 특허문헌의 기술분류로 확인된 목표값을 이용하여 실제 출력값과 목표값의 오차를 이용하는 역전파 알고리즘에 의해 세 개의 은닉층 각각의 가중치가 수정된다. 가중치들은 학습률 η에 따라 확률적 기울기 하강법에 의해 조정된다.

세부기술별 분류는 훈련모집단의 복수의 유효 특허문헌를 이용하여 지도학습을 수행함으로써, 은닉층 내의 가중치가 특정되어 구조화된다. 기술분류 시스템의 테스트 집단에 해당하는 복수의 유효 특허문헌의 기술분류를 수행하여 그 결과를 이용하여 정확도를 파악한다.

3.4 핵심 특허문헌 추출

핵심 특허문헌이란 세부기술별로 분류된 유효 특허문헌 중에서 기술요약서에 기재된 연구개발 기술과 가장 유사하고, 기술적가치가 높은 특허문헌으로서, 특허동향분석 중 정성분석에 필요한 특허문헌이다. 정성분석은 기존의 출원된 특허문헌과 연구개발 기술의 차이에 대한 분석을 수행하는 것이다. 따라서 분석의 목적에 따라 연구개발 기술의 특허출원 또는 실시를 수행하는데 있어 장애가 될 수 있는 장벽특허에 대한 대응방안으로 정보제공, 회피설계, 무효화 방안, 등록된 특허의 권리범위, 공백기술 도출 등을 수행한다.

핵심특허의 추출기준에는 출원된 특허의 유사도, 특허평가, 특허등록여부, 패밀리국가수, 패밀리수, 피인용수, 독립청구항수, 권리만료여부가 있다. 여기서 유사도를 제외한 그 외의 항목은 특허검색서비스에서 검색하여 추출된 후보 특허문헌에 포함된 각 항목에 기재되어 있다.

유사도는 핵심특허를 추출함에 있어, 연구개발 기술과 관련된 특허를 대상으로 하는 것이기 때문에 가장 중요한 항목이다. 다만 유사도는 정형화를 위해 계산된 결과가 0에서 1사이의 값이므로, 다른 선정기준 항목의 값과의 형평성을 위해 유사도 값을 감안하여 10의 제곱승을 곱하여 나타낸다.

특허평가는 특허검색서비스에서 제공하는 값으로, 발명자에 대한 다양한 평가에 의한 발명자 수준, 해당 특허의 기술 영향력, 기술지속성, 시장성, 기술집중도, 경쟁사 견제정도 등에 대하여 산출된 평점이다. 해당 특허에 대하여 산출된 평점은 전체 동종분야 특허와의 순위 산출에 의한 등급으로 1 ~ 10사이의 점수로 평가된다[92].

특허등록여부는 발명자의 발명기술이 특허청에 출원된 후 일정기간이 경과되며 공개되고, 심사관의 심사를 통해 등록이 되면 독점적 권리인 특허권이 발생한다. 공개만 된 발명도 조사대상 특허와 관련성이 있다면 중요하지만, 등록이 된 발명의 경우에는 기술이 새롭고, 향상된 기술임이 검증된 것이므로 중요성이 매우 높아진다.

패밀리국가수는 해당 특허와 동일한 특허가 몇 개의 국가에 출원되었는지를 나타낸다. 패밀리국가수가 많다는 것은 해당 특허기술의 시장성이 높다는 것으로 가치있는 특허를 의미한다.

패밀리수는 해당특허의 국내 또는 국제적인 관련 출원의 정도를 나타낸다. 패밀리 수가 많은 특허는 출원인인 회사나 기관에서 중요하고 기술내용이 가치가 있고, 발명자가 관련 특허를 많이 생산하였을 것으로 판단할 수 있다.

피인용수는 다른 특허들이 해당 특허를 인용한 횟수를 의미한다. 피인용수가 높다는 것은 다른 특허에 의해 많이 인용된 특허이므로 해당 특허의 분야에서 발명의 수준이 높은 특허로 볼 수 있다.

독립청구항은 발명기술을 기본이 되는 청구항으로서, 독립청구항을 근간으로 종속적인 추가 기술이 출원된다. 하나의 특허문헌에 독립청구항의 수가 많다는 것은 관련 기술분야와 관련하여 근원적인 기초기술이 많이 포함되어 있는 것으로 볼 수 있다.

특허권은 특허권을 획득한 발명이 특허출원일로부터 등록 후 20년간 권리를 가지게 되는데, 권리가 만료된 기술은 활용이 가능하므로 연구개발의 대상을 정하는데 참조사항 중에 중요하다. 따라서 핵심 특허문헌을 추출하는 우선순위는 표 3.5와 같이 유사도, 특허평가, 등록여부, 패밀리국가수, 패밀리수, 피인용수, 독립청구항의 수, 권리만료여부의 순이다.

표 3.5 핵심 특허문헌 추출을 위한 항목의 우선순위

Table 3.5 The priority of patent features for extracting core patent documents

	유사도	특허 평가	등록	패밀리 국가수	패밀 리수	피인용 수	대표 청구항	권리만료
우선순위	1	2	3	4	5	6	7	8

그리고 각 항목이 가지는 기본 값을 우선적으로 반영할 수 있도록 우선순위가중치는 1.0 에서 2.0 사이의 값으로 지정하였다. 이러한 핵심 특허문헌의 추출기준이 되는 각 항목의 우선순위에 따른 가중치는 표 3.6과 같이 넷 세트로 각 세트별 우선순위가중치를 적용하여 핵심 특허문헌을 추출한 결과를 비교하고자 한다. 표 3.6의 가중치기호에서 우선순위가중치(w)에 부가된 위첨자는 세트의 번호를 나타내고, 아래첨자는 각 항목을 나타낸다.

표 3.6 우선순위가중치 세트

Table 3.6 The sets of priority weights

가중치 세트	유사도 (w_s^i)	특허 평가 (w_e^i)	등록 (w_r^i)	패밀리 국가수 (w_n^i)	패밀리수 (w_f^i)	피인용수 (w_c^i)	대표 청구항 (w_i^i)	권리 만료 (w_t^i)	항목간 간격
첫 번째 세트	1.77	1.66	1.55	1.44	1.33	1.22	1.11	1.00	0.11
두 번째 세트	1.84	1.72	1.6	1.48	1.36	1.24	1.12	1.00	0.12
세 번째 세트	1.91	1.78	1.65	1.52	1.39	1.26	1.13	1.00	0.13
네 번째 세트	1.98	1.84	1.7	1.56	1.42	1.28	1.14	1.00	0.14

실험 대상은 세부기술별로 분류된 유효 특허문헌을 대상으로 상기 선정된 기준항목에 해당하는 값과 우선순위가중치를 곱하고, 이들의 합산한 선정값은 식 (3.20)과 같다. 여기서, 유사도는 다른 항목의 값이 정수이므로 형평성을 위해 10을 곱하여 나타낸다. 그리고 등록여부와 권리만료여부는 특허권의 유효성의 판단이므로 1과 0의 불리언값으로 표시한다.

$$\text{선정값} = (\text{유사도} \times w_s^i) + (\text{특허평가} \times w_e^i) + (\text{등록} \times w_r^i) + (\text{패밀리국가수} \times w_n^i) \quad (3.20)$$
$$+ (\text{패밀리수} \times w_f^i) + (\text{피인용수} \times w_c^i) + (\text{대표청구항} \times w_i^i) + (\text{권리만료} \times w_t^i)$$

각 유효 특허문헌에 대하여 선정값을 산출하여 내림차순으로 정렬하고 연구 개발기술에서 필요하거나 분석보고서에서 요청된 특허문헌의 수를 선정하여 핵심특허를 추출한다. 그리고 우선순위가중치의 세트를 선정하는데 있어 우선순위가 유사도에서 권리만료의 순이기 때문에 각 세트간의 가중치간격에 따른 우선순위가중치 값의 차이도 표 3.7과 같이, 유사도 항목이 0.07로 가장 크고, 권리만료 항목이 0.00으로 가장 적다.

표 3.7 우선순위가중치 세트간 간격

Table 3.7 The intervals among priority weight sets

가중치 세트	유사도 (w_s^i)	특허 평가 (w_e^i)	등록 (w_r^i)	패밀리 국가수 (w_n^i)	패밀리수 (w_f^i)	피인용수 (w_c^i)	대표 청구항 (w_i^i)	권리 만료 (w_t^i)	간격
세트별 간격	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0	0.01

제 4 장 실험 및 평가

4.1 실험 절차의 개요

본 논문에서는 방사성의약품 이용기술 개발에 관한 특허동향분석을 대상으로 실험하였다. 그림 4.1에서와 같이, 본 실험과 관련한 절차는 다음과 같다.

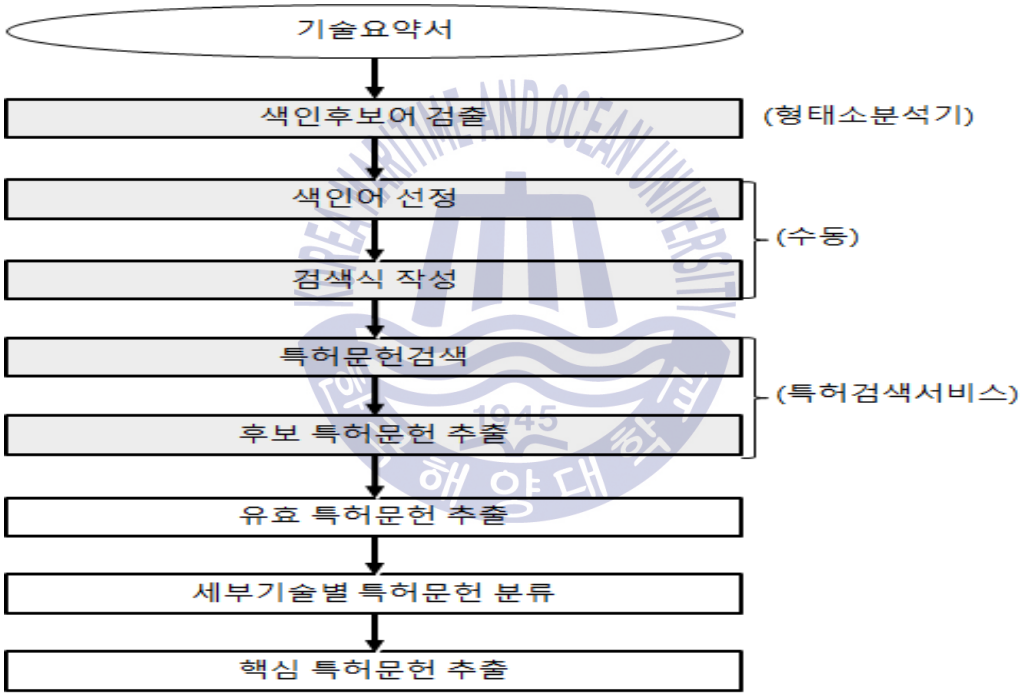


그림 4.1 핵심 특허문헌 추출시스템을 이용한 실험 절차

Fig. 4.1 The experimental procedure for core patent documents extraction system

먼저, 원자력관련 연구소의 연구원이 제안한 기술요약서로부터 검색식을 작성하기 위하여 형태소분석기를 이용하여 기술요약서에 기재된 문장에 대하여 명사 기반의 색인후보어를 검출한다. 검출된 색인후보어 중에서 방사성의약품 이용기술에 관한 색인어를 수동으로 선정한다.

선정된 색인어를 이용하여 검색식의 틀을 작성하고, 색인어의 유사어를 검색하여 한국어 검색식을 작성한다. 그리고 색인어와 검색된 유사어에 대한 영어 단어를 조사하여 한국어 검색식과 동일한 구조로 형성된 영어 검색식을 작성한다. 이러한 검색식의 작성은 수작업을 통하여 이루어진다.

특허검색서비스에서 작성된 검색식을 이용하여 방사성의약품 이용기술 개발과 관련하여 종래 공개되어 있는 특허문헌을 검색하여 특허검색서비스에서 가공되어 있는 후보 특허문헌을 추출한다. 추출받은 후보 특허문헌의 목록은 csv파일 형태로 다운받은 각 후보 특허문헌과 기술요약서간에 유사도를 측정하기 위하여 후보 특허문헌 목록을 가공한다.

가공된 각 특허문헌과 기술요약서의 자질로서 TF-IDF가중치와 항목별가중치, 비색인어가중치를 입력데이터로 문서벡터를 연산하여 각 특허문헌과 기술요약서간의 유사도를 측정한다. 측정된 유사도를 기준으로 후보 특허문헌을 순위화하여 유효 특허문헌을 추출한다.

추출된 유효 특허문헌의 TF-IDF가중치와 기술분류 관련 분류가중치를 이용한 신경망 알고리즘을 활용하여 방사성의약품 이용기술의 하위레벨 세부기술별로 분류한다. 마지막으로, 세부기술별로 분류된 유효 특허문헌 중에서 핵심특허를 추출하는 기준의 항목에 기재된 값과 우선순위가중치를 선형 결합하여 최종적으로 각 세부기술별 핵심특허를 추출한다.

4.2 실험 환경

본 논문에서는 방사성의약품 이용기술 개발에 관한 특허동향분석을 대상으로 실험하였다. 핵심 특허문헌 추출시스템은 자질추출이나 유사도 연산 및 기계학습을 위하여 연산능력이 우수한 별도의 시스템 환경에서 리눅스 데비안 계열의 우분투 버전 16.04를 운영체제로 사용하였다. 그리고 개발 프로그램 언어로 Python 3.5.2를 이용하였다.

기술요약서는 Python 프로그램에 입력으로 사용하기 위해 아래아 한글 파일로 프로그램의 입력을 위해 텍스트 파일로 변환하기 위해 마이크로소프트사 워드 파일로 별도 저장하여 텍스트 변환 프로그램인 "Doc2Txt"를 이용하여 텍스트 파일로 변환한다.

형태소 분석기는 표 4.1과 같이, 한국어 텍스트를 분석할 수 있는 한국어 형태소 분석기를 선택하였으며, 한국어 형태소 분석기 중에서 개발언어인 Python 언어를 지원할 수 있도록 Python 라이브러리를 제공하는 KoNLPy 버전 0.4.3을 이용하였다. 기술요약서를 텍스트 형태로 변환된 파일을 형태소분석기에 입력하여 명사인 단어를 검출하여 색인후 보어 목록을 작성한다. 영어 특허문헌에 대한 텍스트를 형태소 분석하기 위해서는 Python 라이브러리를 제공하는 NLTK를 이용한다.

표 4.1 형태소분석기

Table 4.1 The morphological analyzer

형태소 분석기	분석언어	프로그램 지원 언어
KoNLPy	한국어	Python, Java
NLTK	영어	Python

검색된 후보 특허문헌 목록 및 유효 특허문헌, 핵심 특허문헌은 텍스트 기반의 csv 파일이다. 다만 특허동향분석에서의 차트 이용을 위해 엑셀(.xlsx) 파일로 변경하여 사용된다.

4.3 실험 자료

본 논문에서는 방사성의약품 이용기술 개발에 관한 특허동향분석을 대상으로 실험하였다. 먼저 방사성의약품 이용기술 개발의 기술요약서는 그림 4.2에서와 같이 과제명, 연구개발목표, 연구개발내용, 연구개발 성과, 활용계획 및 기대효과를 포함한다.

양식A201

과제명 : 난치성질환 표적 진단 및 치료 컨버전스 방사성의약품 이용기술 개발			
연구개발목표 (500자 내외)	종양, 퇴행성 뇌질환 등 치료가 어려운 난치성질환에 대하여 방사성동위원소 또는 방사성의약품을 이용한 치료기술 및 치료대상의 선별, 치료효과를 모니터링 할 수 있는 영상진단기술을 융합한 신개념 방사성의약품 및 활용 기술 개발		
연구개발내용 (1000자 내외)	○ 난치암 컨버전스 방사성의약품 안전성 및 유효성 평가 - 분자영상 리포터 발현 유방암 및 전립선암 세포주를 이용한 이종이식 종양 마우스 모델 제조 - 영상을 통한 종양 마우스 모델의 정량 분석 - 종양의 정량 평가를 위한 대사 기반 PET 영상 획득 - 해부학적 영상을 이용한 종양 부피 변화의 정량적 평가 - 컨버전스 방사성 의약품의 분자 표적 영상 획득 및 평가 - 질환 동물모델에서 컨버전스 방사성의약품의 생체분포 분석 - 질환 동물모델에서 컨버전스 방사성의약품의 혈액내 PK 분석 - 컨버전스 방사성의약품의 전임상 단계 안전성 평가를 위한 영상기반 흡수선량평가 - GLP 인증 안전성 평가 ○ 방사선 테라그노시스 개발 물질 최적화 - 종양 이중타겟 방사성 생분해 입자 치료/영상 최적화 - 방사선 나노 테라그노스틱 물질의 생체 분자영상을 통한 최적화 - 방사성표지 치료용 줄기/면역세포 추적 및 치료 최적화		
연구개발성과	○ 난치성 질환 컨버전스 방사성의약품 후보 개발 ○ 영상기반 컨버전스 방사성의약품 정량분석기술 전임상 적용 ○ 난치성질환 치료용 방사성 테라그노시스 임상적용 신약 후보물질 또는 기술 개발 (3종)		
활용계획 및 기대효과 (500자내외) (응용분야 및 활용범위 포함)	○ 난치성 종양 진단/치료 컨버전스 기술 기반 컨버전스 방사성의약품 개발로 적절한 대상 환자군 선정과 표적치료제에 내성이 있는 환자군 치료를 통해 희귀난치성 질환의 치료효과 증대에 기여 ○ 난치성 종양 진단/치료 의약품개발기술은 급성장이 예상되는 바이오시밀러를 포함한 바이오신약 시장 선점이 가능하여 의료 및 유관산업분야 발전 견인 ○ 진단/치료 컨버전스 방사성의약품은 난치성 암질환 극복과 일반의약품의 화학적 독성으로 인해 치료에 어려움이 있는 노약자를 포함한 중증환자 치료에 효과적임.		
중심어_국문	방사성동위원소	난치암	컨버전스 방사성의약품
	질환모델	표적치료	핵의학영상
	마이크로도징		

그림 **4.2** 기술요약서

Fig. 4.2 A technical summary

색인후보어는 기술요약서에 기재되어 있는 서술형의 문장들에 대하여 형태소 분석을 통해 단어의 형태로 검출한다. 따라서 본 논문에서는 기술문헌인 기술요약서에 서술되어 있는 연구개발 기술을 나타내는 문장으로부터 단어들 중에서 명사를 검출하기 위해서 형태소분석기를 사용하였다. 그림 4.3은 형태소 분석기에서 분석할 수 있도록 기술요약서의 한글 파일(hwp)을 텍스트 파일(txt)로 변환하였다.

그림 **4.3** 기술요약서의 텍스트

Fig. 4.3 The text of the technical summary

기술요약서로부터 색인후보어를 검출하기 위하여 먼저 한국어 형태소분석기인 KoNLPy를 이용하여 텍스트 형식의 기술요약서(tech_summary.txt 파일)를 입력하여 기술요약서에 기재되어 있는 여러 문장으로부터 각 단어들의 품사들을 태깅하여 색인후보어로 사용할 명사에 해당하는 단어를 검출하여 색인후보어 파일(indexcandidate.txt)에 저장하였다.

저장된 색인후보어 파일의 색인후보어 목록은 표 4.2와 같다. 색인후보어는 총 131개이다.

표 4.2 기술요약서의 색인후보어 목록

Table 4.2 The list of index candidates in the technical summary

방사성의약품	모델	대사	선별
컨버전스	동물모델	치료대상	줄기세포
치료	마우스	해부학적	암질환
종양	정량	적용	효과적
개발	기반	생체분포	과제명
평가	인증	퇴행성	일반의약품
영상	발현	이종이식	선정
질환	활용계획	연구개발내용	정량분석기술
난치성	적절	이중타겟	의약품
최적화	제조	환자	임상적용
난치성질환	발전	내성	개념
진단	환자군	생체	바이오신약
이용	리포터	화학적	치료기술
기술	바이오시밀러	후보물질	표적치료제
안전성	세포주	면역세포	연구개발성과
방사성	방사성표지	뇌질환	추적
포함	등	독성	선점
분석	활용범위	혈액내	등가균일선량
분자영상	나노입자	급성장	대사종양부피
물질	양전자방출단층촬영	유방암	마이크로도징
내외	기여	흡수선량평가	예상
영상기반	모니터링	정량적	변화
테라그노시스	표적치료	견인	활용
치료효과	생물학적	단계	노약자
치료용	생분해	어려움	융합
항체	전립선암	영상보정	유관산업분야
획득	의약품개발기술	시장	중증환자
방사선	기대효과	수	영상진단
표적	1000자	이용기술	유효선량
방사성동위원소	가능	증대	연구개발목표
500자	후보	응용분야	3종
난치암	신약	질환모델	극복
전임상	희귀난치성	유효성	

표 4.2의 기술요약서에서 검출된 색인후보어 중에서 방사성의약품 이용기술 개발과 관련된 색인어를 선정한다. 색인후보어 중에서 방사성의약품 이용기술 개발과 관련된 색인어를 선정한 결과는 표 4.3과 같다.

표 4.3 기술요약서의 색인어 목록

Table 4.3 The list of indexes in the technical summary

방사성의약품	모델	대사	선별
컨버전스	동물모델	치료대상	줄기세포
치료	마우스	해부학적	암질환
종양	정량	적용	효과적
개발	기반	생체분포	과제명
평가	인증	퇴행성	일반의약품
영상	발현	이종이식	선정
질환	활용계획	연구개발내용	정량분석기술
난치성	적절	이중타겟	의약품
최적화	제조	환자	임상적용
난치성질환	발전	내성	개념
진단	환자군	생체	바이오신약
이용	리포터	화학적	치료기술
기술	바이오시밀러	후보물질	표적치료제
안전성	세포주	면역세포	연구개발성과
방사성	방사성표지	뇌질환	추적
포함	등	독성	선점
분석	활용범위	혈액내	등가균일선량
분자영상	나노입자	급성장	대사종양부피
물질	양전자방출단층촬영	유방암	마이크로도징
내외	기여	흡수선량평가	예상
영상기반	모니터링	정량적	변화
테라그노시스	표적치료	견인	활용
치료효과	생물학적	단계	노약자
치료용	생분해	어려움	융합
항체	전립선암	영상보정	유관산업분야
획득	의약품개발기술	시장	중증환자
방사선	기대효과	수	영상진단
표적	1000자	이용기술	유효선량
방사성동위원소	가능	증대	연구개발목표
500자	후보	응용분야	3종
난치암	신약	질환모델	극복
전임상	희귀난치성	유효성	

불리언 검색은 색인어 단어를 이용하여 AND, OR, NOT과 같은 기본연산자와 NEAR, WITHIN 등과 같은 인접연산자 등의 조합을 통해서 해당 색인어 단어가 포함된 특허문헌을 검색한다. 특허검색서비스를 이용하여 찾고자하는 특허문헌을 검색하기 위하여 검색식을 작성하여야 한다. 일반적으로 검색식에 대한 특정한 양식은 정해져 있지 않다. 하지만 일반적으로 보다 정확한 검색을 위해서는 불리언 검색에 적당한 색인어의 선정 및 연산자조합이 중요하다. 표 4.3의 색인어 중에서 방사성의약품 이용기술에 대한 넓은 범위의 기술용어에서 좁은 범위의 기술용어로 선정된 색인어의 결과는 그림 4.4와 같다.

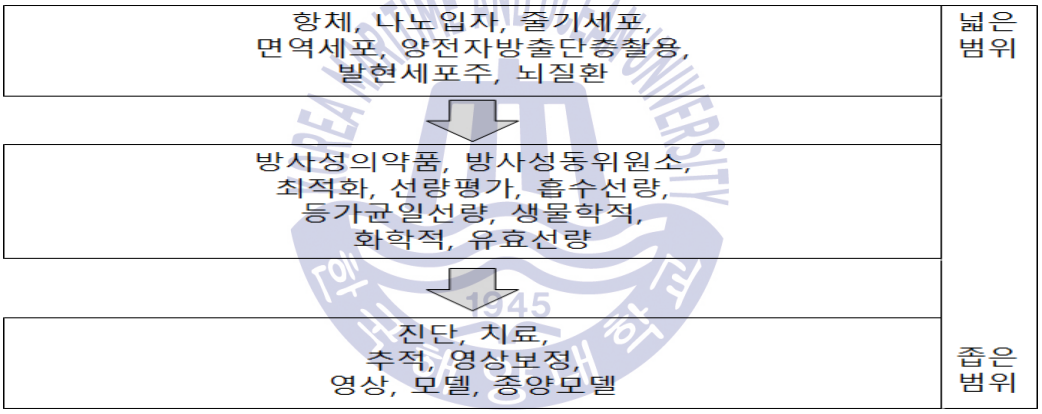


그림 4.4 방사성의약품 이용기술에 대한 기술범위에 기반한 색인어

Fig. 4.4 An index based on the technical scope of radiopharmaceutical technology

선정된 색인어를 기준으로 같은 범위에 속하는 색인어들은 OR 조합을 넓은 범위에서 좁은 범위에 속하는 색인어간에는 AND 조합을 통해 그림 4.5와 같이 검색식의 기본 형태를 작성한다. 여기서 괄호안의 쉼표의 해당 색인어의 유사어와 외래어를 조사하여 OR 조합을 통해 한국어 검색식을 작성한 결과는 그림 4.6과 같다.

(항체 OR 나노입자) AND (방사성의약품 OR 방사성동위원소) AND (진단 OR 치료)

그림 4.5 검색식의 기본형

Fig. 4.5 A template of boolean query

TAC=((항체 OR 펩타이드 OR 저분자화합물 OR 나노입자) AND (방사성의약품 OR "플루오르18" OR "플루오르-18" OR "18플루오르" OR "18-플루오르" OR "플루오린18" OR "플루오린-18" OR "18플루오린" OR "18-플루오린") AND (진단 OR 치료 OR 테라그노스틱)) OR TAC=((줄기세포 OR 면역세포) AND 추적 AND 진단) OR TAC=((PET OR "Positron Emission Tomography" OR SPECT OR "single photon emission computed tomography" OR "양전자 방출 단층 촬영" OR "양전자방출단층촬영" OR "양전전자방출 단층촬영" OR "양전자방출단층촬영" OR "양전자방출, 단층 촬영" OR "단일광자단층촬영" OR "단일광자 단층촬영" OR "단일광자 단층 촬영" OR "단일 광자 단층촬영" OR "단광자방사선단층촬영" OR "단광자방사선 단층촬영" OR "단광자 방사선 단층촬영" OR "단광자방사선 단층 촬영" OR "단광자 방사선 단층 촬영") AND (방사성동위원소 OR "플루오르18" OR "플루오르-18" OR "18플루오르" OR "18-플루오르" OR "플루오린18" OR "플루오린-18" OR "18플루오린" OR "18-플루오린") AND (이벤트 OR 재추출 OR 에너지윈도우 OR 최적화)) OR TAC=((PET OR "Positron Emission Tomography" OR "양전자 방출 단층 촬영" OR "양전자방출단층촬영" OR "양전전자방출 단층촬영" OR "양전자 방출단층촬영" OR "양전자방출 단층 촬영" OR "자기 공명 영상" OR "자기공명 영상" OR "자기 공명영상" OR "자기공명영상") AND 종양 AND (질감분석 OR 이질성 OR 엔트로피 OR 대사종양부피 OR 영역분할)) OR TAC=((PET OR CT OR "Positron Emission Tomography" OR "computed tomography" OR "양전자 방출 단층촬영" OR "양전자방출단층촬영" OR "양전전자방출 단층촬영" OR "양전자 방출단층촬영") OR ((컴퓨터 OR 전산화) AND 단층촬영) AND (선량평가 OR 흡수선량 OR 등가균일선량 OR (생물학적 AND 유효선량))) OR TAC=((유전자 OR 발현세포주) AND 종양모델) OR TAC=(염증 AND 모델 AND (개발 OR 활용)) OR TAC=((뇌질환 OR 뇌출혈 OR 뇌일혈 OR 뇌경색 OR 뇌혈전 OR 뇌색전 OR 뇌동맥 OR 뇌허혈증 OR 뇌염 OR 수막염 OR 간질 OR 파킨슨병 OR 뇌종양 OR 뇌수종 OR 뇌성마비) AND 모델)

그림 4.6 한국어 검색식

Fig. 4.6 The Korean Boolean query

미국, 유럽, 중국 등에 출원된 특허문헌을 조사하기 위해서는 영어 검색식이 필요하다. 작성된 한국어 검색식을 바탕으로 영어단어를 조사하여 영어 검색식을 작성한 결과는 그림 4.7과 같다.

영어 검색식의 경우에는 검색식 내에 '*'를 자주 사용하게 되는데, 이는 영어의 특징인 단어의 품사에 따라 어미가 변화하는 특징을 반영한 결과이다. 검색식에서 '*'는 절단 연산자를 의미하며, 절단 연산자 앞에 입력한 색인어의 후방에 글자 수에 상관없이 추가된 단어까지도 검색할 수 있다.

TAC=((antibody* OR peptid* OR macromolecul* OR nanoparticl*) AND (radiopharmaceuticals OR "radio pharmaceuticals" OR "f18" OR "f-18" OR "18f" OR "18-f") AND (pictur* OR imag* OR vedio* OR monitor*) AND (diagnosis* OR treatment* OR cure* OR remedy* OR therapy* OR care* OR treat* OR theragnostic*)) OR TAC=(((stem* AND cell*) OR immunocyt*) AND (trace* OR chase* OR pursu* OR track*) AND diagnosis*) OR TAC=((PET OR "Positron Emission Tomography" OR SPECT OR "single photon emission computed tomography") AND (radiopharmaceuticals OR "radio pharmaceuticals" OR "f18" OR "f-18" OR "18f" OR "18-f") AND (event* OR resampl* OR "energy window" OR optimiz*)) OR TAC=((PET OR "Positron Emission Tomography" OR MRI OR "Magnetic Resonance Imaging") AND tumor AND ("Texture Features" OR difference OR "heterogeneous nature" OR "metabolic tumor volume" OR entropy OR "area segmentation" OR "region segmentation")) OR TAC=((PET OR CT OR "Positron Emission Tomography" OR "computed tomography") AND ("dose assessment" OR "absorbed dose" OR "Equivalent Dose" OR (biological AND ("effective dose" OR "equivalent dose" OR "equivalent radiation dose")))) OR TAC=(report* AND (gene* OR transfect* OR "cell line") AND tumor* AND model*) OR TAC=(((infect* OR inflammat* OR "inflammatory pain") W/2 model*) AND (develop* OR exploit* OR Applicat* OR use* OR utiliz*)) OR TAC=((cerebropathia* OR cerebropathy* OR "brain desease" OR "cerebral hemorrhage" OR apoplexy OR "cerebral infarction" OR "ischemic stroke" OR "cerebral thrombosis" OR cerebral OR "subarachnoid hemorrhage" OR "cerebral aneurysm" OR stroke OR "brain abscess" OR encephalitis OR meningitis OR epilepsy OR "Parkinson's disease" OR encephaloma OR hydrocephalus OR "cerebral palsy") AND model*)

그림 4.7 영어 검색식

Fig. 4.7 The English Boolean query

특허검색서비스는 최근에 많은 일반적인 검색서비스에서 제공되고 있는 문장검색, 이미지 검색, 음성 검색 등은 제공하지 않으며, 문장검색은 극히 일부 제공하고 있다. 그러나 특허문서 검색을 위한 대부분의 상용시스템은 불리언 검색에 기반을 두고 있다.

이렇게 문서의 유사도나 질의에 따른 순위화를 할 수 없는 불리언 모델을 특허 검색 모델로 하는 이유는 일반문서를 대상으로 하는 기존의 벡터 모델이나 확률 모델이 기존 특허 문서의 데이터베이스에 그대로 적용하기 어렵다. 그리고 하나의 특허 문헌이 하나의 벡터나 수치로 표현하기 어려운 상이한 특징의 텍스트/필드들의 조합이기 때문이다[93].

기술에 관한 문서 중 대표적인 특허문헌은 국가별로 법률의 원칙이나 종래의 관습으로 인하여 특허제도 역시 각 나라별로 제도적인 차이가 있다. 하지만 국제화의 영향으로 세계지적재산권기구에 가입된 국가들간의 특허제도 통일성의 노력으로 특허출원 후 1년 6개월이 경과되거나 특허출원인이 신청한 경우, 특허청에 출원된 특허문헌을 공개하고 있다[94].

공개된 특허문헌에 대한 특허 검색 서비스는 기본적으로 각 국가별 특허청에서 제공하고 있다. 한국 특허청의 경우는 키프리스(KIPRIS, Korea Intellectual Property Rights Information Service)이고, 미국 특허상표국은 USPTO(United States Patent Trademark Office)이며, 일본 특허청은 JPO(Japan Patent Office), 유럽 특허청에는 이스페이스넷(espacenet)이 무료로 제공되고 있다.

하지만 상기 특허청에서 운영되는 특허검색서비스는 자국의 특허문헌을 중심으로 제공하고 있다. 그리고 다른 국가의 특허문헌에 대하여 별도의 가공이 없어 검색하는데 있어 어려움과 전체적인 검색 시간 및 검색 결과의 다운로딩 시간, 연산자 개수의 제한 등의 문제가 있다.

이에 일반적으로 특허동향분석에서는 한국 뿐 아니라 미국, 일본, 유럽, 중국 등 여러 국가의 특허문헌을 같이 검색할 수 있는 유료 특허검색서비스를 많이 사용하고 있다. 유료의 특허검색서비스의 종류에는 ㈜윕스의 윕스온과 윈텔립스, ㈜위즈도메인의 포커스트, ㈜톰슨 로이터의 톰슨 등이 있다. 본 논문에서는 ㈜위즈도메인의 포커스트를 사용한다.

이러한 유료 또는 무료의 특허검색서비스에서 검색방법은 특허문헌의 서지사항과 관계되는 출원번호 또는 공개번호 등을 이용한 번호검색, 출원인 또는 발명자 등을 이용한 인명검색, IPC코드에 의한 검색, 키워드 검색을 제공한다. 그 중에서도 일반적으로 많이 사용되는 검색방법은 특허조사원들이 색인어를 이용하여 해당 기술과 관련된 특허문헌을 검색하는 방법인 키워드 검색이다.

그림 4.8은 기술요약서로부터 작성된 검색식에 의해 특허검색서비스인 ㈜위즈도메인의 포커스트를 이용한 후보 특허문헌의 검색화면이다. 그림 4.9는 검색된 결과를 다운받은 csv형태의 파일이다.

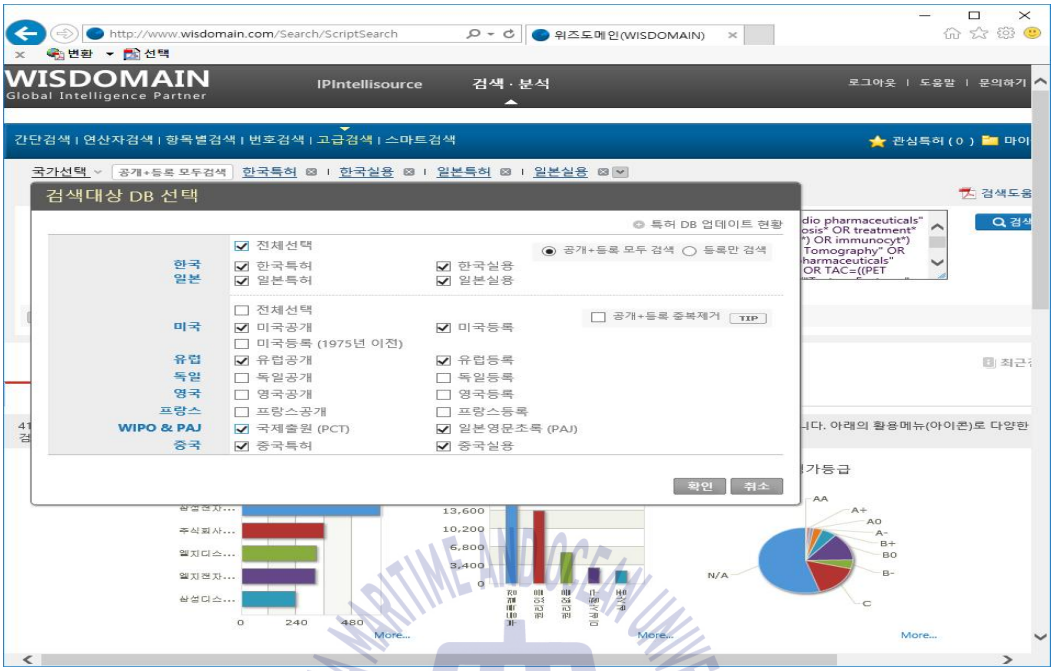


그림 4.8 후보 특허문헌의 검색 화면

Fig. 4.8 The search screen shot of candidate patent documents

일련번호	번호	명칭	명칭(원문)	요약	요약(원문)	출원인	출원인영문	출원인국기	출원인대표	발명자	국제특허분	공통특허분	미국특허분
1	EP0230889	Internalizing anti-CD7		The present invention		Immunomedics, Inc.		US		Hansen, Ha	C07K-016/28,A61K-039/395,A61P		
2	US200903(	Anti-Pancreatic Cance		Described herein are		IMMUNOMEDICS, INC.		US	IMMUNOMI	Goldenberg	A61K-039/	A61K47/48	424/001.4
3	US200602	Stably tethered structu		The present invention		IBC PHARMACEUTICALS, INC.			IBC PHARM	Chang; Chi	A61K-051/	C07K16/18	424/001.4
4	US2005001	Monoclonal antibody h		This invention relates		Immunomedics, Inc.			IMMUNOMI	Goldenberg	G01N-033/	A61K49/00	435/007.2
5	JP2012509	IL-6に 対 ヒ Ｉ Ｌ - 6 に		(57)【要約】	(57)【要約】	アルダ- バイオファ-		US		ガルシア,	A61K-039/	C07K16/248,A61K203	
6	EP0151995	HUMANIZED MONOCL		This invention relates		Immunomedics, Inc.		US		GOLDENBE	C07K-016/	A61K49/0008,A61K47/	

그림 4.9 검색된 후보 특허문헌

Fig. 4.9 The retrieved candidate patent documents

특허검색서비스를 통해 4306건의 후보 특허문헌을 획득하여 실험에 활용하였다. 검색된 후보 특허문헌 목록은 해당 색인어가 포함된 모든 특허문헌을 검색하기 때문에 검색식에 포함된 색인어를 내포하고 있다. 하지만 언어의 특징인 단어에 대한 동음이의어가 존재할 수 있고 동일한 단어를 이용하여 다른 기술을 묘사할 수 있으며 전체 문장이 다른 의미를 나타내는 등의 표현상의 차이로 인한 문제점이 있어, 실제 기술요약서에 표시된 연구개발 기술과 전현 연관성이 없는 특허문헌도 포함되는 불리언 검색의 한계가 나타난다.

4.4 평가 방법

핵심 특허문헌 추출시스템은 특허동향분석의 절차 중 특허검색서비스를 통하여 추출한 후보 특허문헌 목록을 대상으로 유사도 랭킹 영역을 통한 유효 특허문헌의 추출, 분류확률 랭킹 영역을 통한 세부기술별 분류, 선정값 랭킹에서의 핵심 특허문헌 추출을 시스템화하여 수행하였다. 본 논문에서는 추출시스템을 통하여 수행한 결과를 특허조사원이 수작업을 통하여 수행한 결과에 대한 정확률과 각 절차를 수행하는데 요구되는 시간을 기준으로 비교하여 성능을 평가하고자 한다. 특허조사원은 화학 관련 분야를 전공하였고, 특허동향분석 업무에 있어 5년 이상의 경력을 가지고 있다.

먼저 유효 특허문헌의 추출과 관련하여 특허조사원이 수작업으로 수행한 결과를 기준으로 복수의 항목별가중치 세트를 적용하여 추출된 유효 특허문헌을 대상으로 정확률에 관한 성능을 비교하여 최적의 항목별가중치 세트를 선택한다. 그리고 유효 특허문헌의 추출절차를 수행하는 데 특허조사원의 수작업과 추출시스템의 자동화 과정간의 수행시간을 측정하여 성능을 평가한다. 그리고 세부기술별 분류에 있어서는 방사성의약품 이용기술에 해당하는 1613건의 유효 특허문헌을 대상으로 제안된 신경망 구조를 이용하여 방사성의약품 개발, 진단 및 치료 성능 평가 기술, 질환모델 생산 기술로 분류하는데 추출시스템을 통해 수행한 시간과 특허조사원이 수작업으로 분류한 시간을 비교하여 추출시스템의 성능을 평가한다.

특허조사원이 분류한 결과와 추출시스템을 통하여 분류된 결과를 비교하여 추출시스템의 정확률을 평가하며, 나아가 다른 특허문헌 분류 알고리즘의 정확률을 비교하여 성능을 평가한다. 마지막으로 핵심 특허문헌의 추출과 관련하여 특허조사원이 수작업을 통해 수행하는데 필요한 시간과 추출시스템의 선정값 연산을 통해 수행한 시간을 비교하여 추출시스템의 성능을 평가한다. 그리고 표 3.6의 우선순위가중치 세트를 각각 적용하여 추출된 핵심 특허문헌의 수와 특허조사원이 선정한 핵심 특허간의 정확률을 비교평가하여 최적의 우선순위가중치 세트를 선정한다.

4.5 성능 평가

4.5.1 유효 특허문헌 추출

특허동향분석에 있어 특허검색서비스에서 검색한 후보 특허문헌 목록으로부터 방사성 의약품 이용기술 개발과 관련된 특허문헌을 추출하기 위해서 기술요약서와 각 특허문헌의 유사도를 측정하기 위해서 먼저 기술요약서의 자질을 추출한다. 기술요약서의 자질은 기술요약서의 문서벡터로서 기술요약서에 기재되어 있는 문장 내 색인후보어들의 TF-IDF가중치이다.

표 4.4는 기술요약서의 각 색인후보어의 TF값이다. 그리고 IDF값은 기술요약서가 하나의 문서로 이루어져 있으므로 1이다. 따라서 기술요약서의 자질인 TF-IDF가중치는 TF값과 같다.

표 4.4 기술요약서내 색인후보어들의 TF값 목록

Table 4.4 The TF values of index candidates in the technical summary

색인후보어	TF값	색인후보어	TF값	색인후보어	TF값
방사성의약품	0.052863	환자군	0.004405	견인	0.004405
컨버전스	0.052863	리포터	0.004405	단계	0.004405
치료	0.044053	바이오시밀러	0.004405	어려움	0.004405
종양	0.030837	세포주	0.004405	영상보정	0.004405
개발	0.026432	방사성표지	0.004405	시장	0.004405
평가	0.026432	등	0.004405	수	0.004405
영상	0.022026	활용범위	0.004405	이용기술	0.004405
질환	0.017621	나노입자	0.004405	증대	0.004405
난치성	0.017621	양전자방출단층촬영	0.004405	응용분야	0.004405
최적화	0.017621	기여	0.004405	질환모델	0.004405
난치성질환	0.013216	모니터링	0.004405	유효성	0.004405
진단	0.013216	표적치료	0.004405	선별	0.004405
이용	0.013216	생물학적	0.004405	줄기세포	0.004405
기술	0.013216	생분해	0.004405	암질환	0.004405
안전성	0.013216	전립선암	0.004405	효과적	0.004405
방사성	0.013216	의약품개발기술	0.004405	과제명	0.004405
포함	0.013216	기대효과	0.004405	일반의약품	0.004405
분석	0.013216	1000자	0.004405	선정	0.004405
분자영상	0.008811	가능	0.004405	정량분석기술	0.004405
물질	0.008811	후보	0.004405	의약품	0.004405
내외	0.008811	신약	0.004405	임상적용	0.004405
영상기반	0.008811	희귀난치성	0.004405	개념	0.004405
테라그노시스	0.008811	대사	0.004405	바이오신약	0.004405
치료효과	0.008811	치료대상	0.004405	치료기술	0.004405
치료용	0.008811	해부학적	0.004405	표적치료제	0.004405
항체	0.008811	적용	0.004405	연구개발성과	0.004405
획득	0.008811	생체분포	0.004405	추적	0.004405
방사선	0.008811	퇴행성	0.004405	선점	0.004405
표적	0.008811	이종이식	0.004405	등가균일선량	0.004405
방사성동위원소	0.008811	연구개발내용	0.004405	대사종양부피	0.004405
500자	0.008811	이중타겟	0.004405	마이크로도징	0.004405
난치암	0.008811	환자	0.004405	예상	0.004405
전임상	0.008811	내성	0.004405	변화	0.004405
모델	0.008811	생체	0.004405	활용	0.004405
동물모델	0.008811	화학적	0.004405	노약자	0.004405
마우스	0.008811	후보물질	0.004405	융합	0.004405
정량	0.008811	면역세포	0.004405	유관산업분야	0.004405
기반	0.008811	뇌질환	0.004405	중증환자	0.004405
인증	0.004405	독성	0.004405	영상진단	0.004405
발현	0.004405	혈액내	0.004405	유효선량	0.004405
활용계획	0.004405	급성장	0.004405	연구개발목표	0.004405
적절	0.004405	유방암	0.004405	3종	0.004405
제조	0.004405	흡수선량평가	0.004405	극복	0.004405
발전	0.004405	정량적	0.004405		

기술요약서의 문서벡터(d^t)는 기술요약서에 기재되어 있는 각 색인후보어의 TF-IDF가 중치의 모임이다. 따라서 식 (4.1)과 같이 표현된다. 기술요약서의 문서벡터(d^t)와 유사도 측정을 위해 방사성의약품 이용기술 개발과 관련하여 검색된 각 후보 특허문헌의 문서벡터(d^p)를 구한다.

$$d_j^t = (w_{\text{방사선의약품},j}^t, w_{\text{컨버전스},j}^t, \cdots, w_{\text{극복},j}^t) = (0.052863, 0.052863, \cdots, 0.004405) \qquad (4.1)$$

그림 4.10은 각 특허문헌의 자질을 측정함에 있어, 시스템의 처리속도를 향상시키기 위해서 후보 특허문헌 목록을 가공하였다. 후보 특허문헌 목록의 각 열에 기재된 특허문헌에서 텍스트 기반의 정보인 발명의 명칭, 요약, 대표청구항, 청구항 항목과 각 특허문헌의 구분자로서, 출원번호를 제외한 나머지 서지사항 등은 제거한다.

번호	발명의 명칭	요약	대표 청구항	전체 청구항
KR2013011	[18 F]플루오르메틸기가 도입	본 발명은 [18F]플루오르메	1Normethyl-PBR28에	1Normethyl-PBR28에 트리아조
KR2013007	14-3-3 시그마 유전자 결손 마	본 발명은 14-3-3 시그마 유	114-3-3 시그마 유전자	114-3-3 시그마 유전자를 결손
KR2012014	2차원 전리함 선량계를 이용한	본 발명은 2차원 전리함 선	12차원 전리함 선량계	12차원 전리함 선량계를 이용한
KR2015018	3,3'-다이인돌일메탄, 이의 전	본 발명은 3,3'-다이인돌일	13,3'-다이인돌일메탄,	13,3'-다이인돌일메탄, 이의 전
KR2005701	3중 작용 시약에 의해 연결된	본 발명은 항림프종 항체어	1 항림프종 항체 또는	1 항림프종 항체 또는 이의 변
KR2005701	3중 작용 시약에 의해 연결된	본 발명은 항림프종 항체어	1 항림프종 항체 또는	1 항림프종 항체 또는 이의 변
KR2012701	I L -6에 대한 항체 및 이들의	본 발명은 질병을 예방하거	1IL-6과 연관된 질환 또	1IL-6과 연관된 질환 또는 이상
KR2012701	I L -6에 대한 항체 및 이들의	본 발명은 질병을 예방하거	1IL-6과 연관된 질환 또	1IL-6과 연관된 질환 또는 이상
KR2014006	A E G - 1 을 유효성분으로 포	본 발명은 AEG-1(astrocyte	1AEG-1(astrocyte elev	1AEG-1(astrocyte elevated ger
JP2008540	AKAP-PKA 상호 작용의 비펩티	(57)【요약】 본 발명은, 담	【청구항 1】 표 A에 의	【특허청구의 범위】【청구항 1】
KR2015006	Atg7+/--ob/ob 형질을 나타내	본 발명은 당뇨병 동물모델	1다음 단계를 포함하는	1다음 단계를 포함하는 당뇨병
JP2009501	atheroma 동맥경화성심 혈관	(57)【요약】 본 발명은, ath	【청구항 1】 이하의 단	【특허청구의 범위】【청구항 1】
JP2013521	Aβ을 표적이라고 하는 면역 요	(57)【요약】【과제】본 발명은	【청구항 1】 Aβ을 표적	【특허청구의 범위】【청구항 1】
JP2013521	Aβ을 표적이라고 하는 면역 요	(57)【요약】【과제】본 발명은	【청구항 1】 Aβ을 표적	【특허청구의 범위】【청구항 1】

그림 4.10 후보 특허문헌 목록에서의 기술내용 항목

Fig. 4.10 The technical fields of candidate patent document list

유사도를 측정하기 위한 각 특허문헌의 자질은 각 특허문헌의 텍스트 정보에 포함되어 있는 색인후보어가 포함된 각 항목별 가중치와 TF-IDF가중치 및 비색인어 가중치의 선형 조합에 의한 특허문헌의 가중치($w_{i,j}^{p}$)이다. 예를 들어, 표 4.5는 특허문헌의 문서벡터 값의 예시를 나타낸다.

표 4.5 특허문헌의 색인후보어별 가중치

Table 4.5 The weights by index candidates of patent documents

색인후보어	$tf_{i,j}^{p}$	idf_{i}^{p}	$tfidf_{i,j}^{p}$	w_{i}^{n}	$w_{i,j}^{p}$
방사선의약품	0.00610	1.736447	0.010604	1.0	0.010604
컨버전스	0.02748	1.885886	0.051826	1.5	0.077739
⋮	⋮	⋮	⋮	⋮	⋮
극복	0.00305	2.935104	0.008962	1.5	0.013443

첫 번째 특허문헌의 첫 번째 색인후보어인 방사선의약품의 경우, 대상 특허문헌의 요약에만 1회 기재되었고, 전체 색인후보어의 수는 131개이므로, 먼저 항목별가중치는 각 특허문헌에 포함된 발명의 명칭, 요약, 대표청구항, 청구항 항목별로 색인색인후보어가 기재된 TF값에 대하여 표 3.2에 기재된 항목별가중치 여섯 세트를 각각 적용하여 식(4.2)를 이용하여 $ttf_{\text{방사선의약품},1}^{p}$값을 구한다. 다만 예시의 경우에는 항목별가중치 네 번째 세트를 적용하였다.

$$ttf_{\text{방사선의약품},1}^{p} = \frac{(f_{\text{방사선의약품},1}^{t} \times \alpha + f_{\text{방사선의약품},1}^{a} \times \beta + f_{\text{방사선의약품},1}^{f} \times \gamma + f_{\text{방사선의약품},1}^{c} \times \delta)}{\sum_{i=1}^{n} f_{i,j}^{p}} \quad (4.2)$$

$f_{\text{방사선의약품},1}^{a}$값은 1이고, 요약에 기재되었으므로 β값인 0.8을 곱하고, 분모는 색인후보어의 수이므로 131을 적용하면, $ttf_{\text{방사선의약품},1}^{p}$값은 0.00610이다. 그리고 IDF값은 항목별가중치의 영향이 없으므로 전체 후보 특허문헌의 건수는 총 4306건 즉, 문서집합의 총 개수인 $|D^{p}|$는 4306이고, 특허문헌 4306건 중의 79건에만 기재되어 1.736447이다.

TF-IDF가중치는 식 (4.3)을 통하여 나온 결과는 0.010604이다. 마지막으로 비색인어가중치는 색인후보어 중에서 검색식에 포함되지 않았던 비색인어의 경우 식 (4.4)를 이용하여 최종 가중치를 계산하고, 색인후보어중 색인어의 경우에는 식 (4.5)를 이용하여 최종가중치를 계산한다.

$$w^{p}_{\text{방사선의약품},1} = tfidf^{p}_{\text{방사선의약품},j} = 0.00610 \times 1.736447 \quad (4.3)$$

$$w^{p}_{\text{방사선의약품},1} = w^{p}_{\text{방사선의약품},1} \times 1.0 \quad (4.4)$$

$$w^{p}_{\text{컨버전스},1} = w^{p}_{\text{컨버전스},1} \times 1.5 \quad (4.5)$$

따라서 색인후보어 방사선의약품은 색인어이므로 비색인어 가중치가 1.0이 적용되어 결과적으로 색인후보어 방사선의약품의 문서벡터는 0.010604이다. 색인후보어 컨버전스는 색인어가 아니므로 비색인어가중치 1.5가 적용되어 0.077739이다.

각 색인후보어에 대하여 계산된 최종가중치를 통해 특허문헌에 대한 문서벡터를 계산한다. 이와 같이 첫 후보 특허문헌에 대한 문서벡터(d^{p}_{1})는 식 (4.6)과 같이, 색인후보어가 특허문헌에 나타나는 TF-IDF가중치이다. 이와 같이 각 후보 특허문헌의 문서벡터(d^{p})들을 구한다.

$$d^{p}_{1} = (w^{p}_{\text{방사선의약품},1}, w^{p}_{\text{컨버전스},1}, \cdots, w^{p}_{\text{극복},1}) = (0.010604, 0.077739, \cdots, 0.013443) \quad (4.6)$$

기술요약서의 문서벡터(d^t)와 각 후보 특허문헌의 문서벡터(d^p)간의 유사도를 측정한다. 유사도 측정방법은 식 (3.12)를 이용하여 구한 기술요약서와 각 특허문헌간의 유사도 측정결과는 그림 4.11과 같다. 유사도 측정값을 후보 특허문헌 목록의 각 열에 새로운 항목인 유사도 항목을 추가하여 입력된다. 입력된 유사도에 따라 내림차순으로 순위화하여 유사도가 높은 순으로 유효 특허문헌을 추출한다.

번호	발명의 명칭	요약	대표 청구항	전체 청구항	유사도
JP2009501318T	atheroma 동맥경화성심 혈관	(57)【요약】 본 발명은, ath	【청구항 1】 이하의 단	【특허청구의 범위】【청구항 1】	0.568328
KR20050039405A	ＦＧＦ2를 유효성분으로 포함	본 발명은 섬유모세포 성장	7 FGF2(Fibroblast Gro	1<AmendStatus status="D">식	0.555279
JP2013521233T	Aβ을 표적이라고 하는 면역 요	(57)【요약】【과제】 본 발명	【청구항 1】 Aβ을 표적	【특허청구의 범위】【청구항 1】	0.504223
KR20140062961A	ＡＥＧ－１을 유효성분으로 포	본 발명은 AEG-1(astrocyte	1AEG-1(astrocyte elev	1AEG-1(astrocyte elevated ger	0.455189
KR20127016163A	ＩＬ-6에 대한 항체 및 이들의	본 발명은 질병을 예방하거	1IL-6과 연관된 질환 또	1IL-6과 연관된 질환 또는 이상	0.378876
KR20120142532A	2차원 전리함 선량계를 이용한	본 발명은 2차원 전리함 선	12차원 전리함 선량계	12차원 전리함 선량계를 이용한	0.337009
KR20057010811A	3중 작용 시약에 의해 연결된	본 발명은 항림프종 항체에	1 항림프종 항체 또는	1 항림프종 항체 또는 이의 변	0.335669
KR20150061829A	Atg7+/--ob/ob 형질을 나타내	본 발명은 당뇨병 동물모델	1다음 단계를 포함하는	1다음 단계를 포함하는 당뇨병	0.319542
KR20067012885A	ＥＲＢ 항원의 표적화(TARGE	a) 삼작용기성 가교결합 부	1 a) 삼작용기성 가교결	1 a) 삼작용기성 가교결합 부위	0.312025
JP2008540586T	AKAP-PKA 상호 작용의 비펩티	(57)【요약】 본 발명은, 담	【청구항 1】 표 A에 의	【특허청구의 범위】【청구항 1】	0.241207
KR20057010811A	3중 작용 시약에 의해 연결된	본 발명은 항림프종 항체에	1 항림프종 항체 또는	1 항림프종 항체 또는 이의 변	0.234437
KR20090063996A	ＦＧＦ2를 유효성분으로 포함	본 발명은 섬유모세포 성장	1 FGF2(Fibroblast Gro	1 FGF2(Fibroblast Growth Fact	0.210024
KR20150181902A	3,3'-다이인돌일메탄, 이의 전	본 발명은 3,3'-다이인돌일	13,3'-다이인돌일메탄,	13,3'-다이인돌일메탄, 이의 전	0.197441
JP2007512324T	ERB 항원의 타겟팅	(57)【요약】 하기 a)~d)을	【청구항 1】 하기 a)~d	【특허청구의 범위】【청구항 1】	0.192844
KR20130110282A	[18Ｆ]플루오르메틸기가 도입	본 발명은 [18F]플루오르메	1Normethyl-PBR28에	1Normethyl-PBR28에 트리아조	0.176844
KR20127016163A	ＩＬ-6에 대한 항체 및 이들의	본 발명은 질병을 예방하거	1IL-6과 연관된 질환 또	1IL-6과 연관된 질환 또는 이상	0.146959
KR20097001229A	ＦＺＤ10에 대한 종양-표적화	본 발명은 프리즐드 상동체	1 서열 15, 17 및 19에	1 서열 15, 17 및 19에 제시된	0.146034
KR20127033515A	ＢＡＭＢＡＭ：고처리율 서열	본 발명은 임상 병상, 예컨	1자등 유전자 서열 대상	1자등 유전자 서열 대상을 유도	0.114218
JP2007537245T	FGF2을 유효 성분으로서 포함	(57)【요약】 본 발명은, 선	【청구항 1】 FGF2을 유	【특허청구의 범위】【청구항 1】	0.074696
KR20130074658A	14-3-3 시그마 유전자 결손 마	본 발명은 14-3-3 시그마 유	114-3-3 시그마 유전자	114-3-3 시그마 유전자를 결손	0.073378
JP2005534641T	FAS 펩티드 모방 몸 및 그 사용	(57)【요약】 Fas을 무능히 유	【청구항 1】 FasL과 상	【특허청구의 범위】【청구항 1】	0.070996
KR20070096662A	ＦＧＦ2를 유효성분으로 포함	본 발명은 섬유모세포 성장	1 FGF2(Fibroblast Gro	1 FGF2(Fibroblast Growth Fact	0.062795
KR20097001229A	ＦＺＤ10에 대한 종양-표적화	본 발명은 프리즐드 상동체	1 서열 15, 17 및 19에	1 서열 15, 17 및 19에 제시된	0.039318
JP2004113791A	CT 스캔에 있어서 자동 노출 제	(57)【요약】【과제】 CT 스캔	【청구항 1】 조점으로부	【특허청구의 범위】【청구항 1】	0.032295
JP2007512324T	ERB 항원의 타겟팅	(57)【요약】 하기 a)~d)을	【청구항 1】 하기 a)~d	【특허청구의 범위】【청구항 1】	0.027936
JP2013521233T	Aβ을 표적이라고 하는 면역 요	(57)【요약】【과제】 본 발명	【청구항 1】 Aβ을 표적	【특허청구의 범위】【청구항 1】	0.027107

그림 4.11 기술요약서와 특허문헌간의 유사도

Fig. 4.11 The similarity between technical summary and patent documents

특허동향분석에서 유효 특허문헌의 범위를 정하는 기준은 연구개발 기술과의 관련성을 기준으로 판단한다. 연구개발 기술의 관련성은 색인어 기반의 검색식을 통해 검색된 후보 특허문헌 목록의 수가 많기 때문에, 특허문헌의 발명의 명칭, 요약 또는 도면을 참조하여 빠르게 읽고 내용을 파악하게 된다.

본 논문에서는 측정된 유효 특허문헌의 유사도를 기준으로 내림차순으로 정렬하여 순위화한 후, 유사도 측정값의 변화를 확인한 결과 급격한 차이가 발생하는 부분을 기준으로 유효 특허문헌의 범위를 추출하였다. 표 3.2의 항목별가중치 여섯 세트를 각각 적용하여 실험한 결과와 수작업을 통하여 얻어진 결과와 비교하였다.

먼저 최적의 항목별가중치를 선정하기 위하여 각 세트를 적용하여 유효 특허문헌의 수를 결정하는데 있어 특허조사원의 수작업에 의해 추출된 유효 특허문헌과의 정확률을 판단하는데 있어 추출된 1613건과 동일한 특허문헌수를 대상으로 하려고 하였다. 그러나 정확률을 판단하는데 있어 추출시스템을 통하여 추출된 유효 특허문헌의 수를 기준으로 특허조사원이 수작업을 통해 추출한 유효 특허문헌과 비교하여 일치하는 특허문헌의 수로 판단하는데 몇 가지 문제점이 있었다.

첫 번째 문제점은 특허조사원이 수작업을 통해 추출한 유효 특허문헌의 수를 기준으로 하게 되면 추출시스템을 통하여 추출된 유효 특허문헌의 수가 많으면 많을수록 정확률이 높아지고, 상대적으로 불일치하는 유효 특허문헌의 수가 많음에도 불구하고 정확률이 높은 문제가 있기 때문이다. 그리고 두 번째 문제점은 세트별 항목별가중치를 이용한 유사도 값에 있어 1613번째 건으로 추출하기에는 기준이 되는 유사도의 값이 너무 비슷하였으며, 추출기준이 되는 특허문헌이 패밀리특허인 경우가 많아 내용적으로 동일하거나 유사한 특허를 함께 포함하거나 제거하는 것이 정확률 판단에 부합한다고 판단하였다.

그림 4.12는 항목별가중치 세트별로 적용한 추출시스템을 통해 추출된 유효 특허문헌의 수를 나타낸다. 그리고 특허조사원의 수작업을 통하여 추출한 1613건의 유효 특허문헌을 기준으로 추출시스템을 통해 추출된 유효 특허문헌의 수와의 차이도 나타내고 있다.

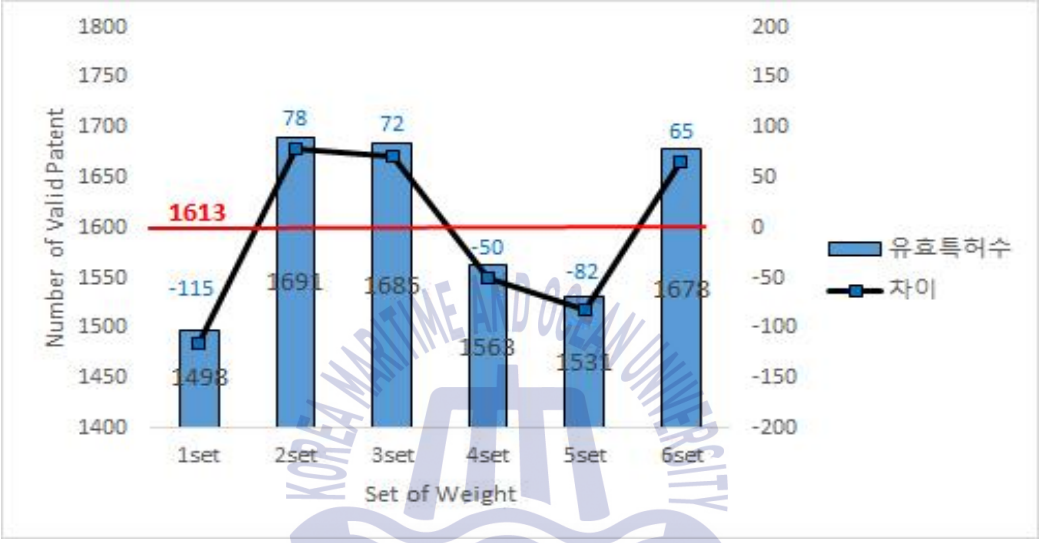


그림 4.12 항목별가중치 세트별 유효 특허문헌의 수

Fig. 4.12 The number of valid patent documents by sets of field weight

먼저 특허조사원이 수작업을 통해 유효 특허문헌을 추출한 결과와 첫 번째 세트의 항목별가중치를 적용한 추출시스템을 통한 유효 특허문헌을 결정한 결과가 표 4.6이다. 그리고 국가별로 추출된 유효 특허문헌의 결과와 전체 유효 특허문헌의 결과를 같이 비교하였다.

국가별로는 한국의 경우, 754건의 유사 특허문헌 목록에서 266건을 수작업으로 추출하여 35.28%인데 반해, 추출시스템의 경우 263건으로 34.88%에 해당한다. 일본은 939건의 유사 특허문헌에서 수작업으로 355건을 추출하여 37.81%인데 반해 추출시스템은 318건으로 33.87%의 비율이다.

표 4.6 첫 번째 항목별가중치 세트 적용 유효 특허문헌 수

Table 4.6 The number of valid patent documents applied 1st field weight set

국가	후보 특허문헌 수	수작업		첫 번째 세트 항목별가중치	
		유효특헌	비율	유효특헌	비율
한국	754	266	35.28%	263	34.88%
일본	939	355	37.81%	318	33.87%
미국	1580	583	36.90%	574	36.33%
유럽	620	257	41.45%	198	31.94%
PCT	413	152	36.80%	145	35.11%
합계	4306	1613	37.46%	1498	34.79%

미국은 1580건의 유사 특허문헌 목록에서 수작업은 583건으로 36.90%의 비율로 추출하였으나, 추출시스템의 경우 574건으로 36.33%의 비율이며, 유럽은 유사 특허문헌 620건 중에서 수작업은 257건으로 41.45%비율이나 추출시스템의 경우에는 198건으로 31.94%이다. 마지막으로 PCT는 413건의 유효 특허문헌 목록에서 152건을 수작업으로 추출하여 36.80%이지만 추출시스템은 145건을 추출하여 35.11% 비율이다.

전체적으로는 특허조사원이 수작업을 통하여 4306건의 유사 특허문헌 목록에서 1613건의 유효 특허문헌을 추출하여 그 비율이 37.46%인데 반해, 추출시스템을 통한 유효 특허문헌을 결정한 결과는 1498건의 유효 특허문헌을 추출하여 34.79%의 비율로 수작업을 이용한 경우보다 115건이 적다. 다만 유럽의 경우, 수작업이 비교적 높은 비율로 유효 특허문헌을 추출하였으나, 추출시스템은 그 비율이 낮아 전체적으로는 추출된 유효 특허문헌의 비율에 영향을 미치고 있다.

표 4.7은 특허조사원이 수작업을 통해 유효특허를 추출한 결과와 두 번째 항목별가중치 세트를 추출시스템에 적용하여 유효 특허문헌을 결정한 결과를 나타낸다. 그리고 특허조사원이 수작업을 통해 추출된 유효 특허문헌의 결과는 첫 번째 세트와 동일하므로 앞으로 추출시스템을 통한 유효 특허문헌의 결과만 언급한다.

한국은 754건의 유사 특허문헌 목록에서 295건으로 39.12%에 해당하며, 일본은 939건의 유사 특허문헌에서 346건으로 36.85%의 비율이고, 미국은 1580건의 유사 특허문헌 목록에서 633건으로 40.06%의 비율이며, 유럽은 유사 특허문헌 620건 중에서 255건으로 41.13%이고, PCT는 413건의 유효 특허문헌 목록에서 162건을 추출하여 39.23% 비율이다.

표 4.7 두 번째 항목별가중치 세트 적용 유효 특허문헌 수

Table 4.7 The number of valid patent documents applied 2nd field weight set

국가	후보 특허문헌 수	수작업		두 번째 세트 항목별가중치	
		유효특헌	비율	유효특헌	비율
한국	754	266	35.28%	295	39.12%
일본	939	355	37.81%	346	36.85%
미국	1580	583	36.90%	633	40.06%
유럽	620	257	41.45%	255	41.13%
PCT	413	152	36.80%	162	39.23%
합계	4306	1613	37.46%	1691	39.27%

전체 후보 특허문헌 목록 4306건 중에서 추출시스템을 통한 유효 특허문헌을 결정한 결과는 1691건의 유효 특허문헌을 추출하여 39.27%의 비율로 수작업을 이용한 경우보다 78건이 많았다. 여기서 첫 번째 세트의 경우와 비교하여 항목별가중치의 조그만 변화에 결과의 차이가 비교적 크게 나타나는 것은 기술내용이 유사한 패밀리특허가 많은 특허에 항목별가중치가 영향을 많이 준 것으로 파악된다.

다음으로 표 4.8은 특허조사원이 유효특허를 수작업으로 추출한 결과와 유효 특허문헌을 결정하는데 있어 세 번째 항목별가중치 세트를 추출시스템에 적용한 결과를 보여준다. 한국의 경우는 754건의 유사 특허문헌 목록 중에서 288건으로 38.20%에 해당하며, 일본의 경우는 939건의 유사 특허문헌 중에서 36.74%의 비율인 345건이고, 미국은 1580건의 유사 특허문헌 목록에서 624건으로 39.49%의 비율이며, 유럽의 경우는 유사 특허문헌 620건에서 261건으로 42.10%이다. PCT는 413건의 유효 특허문헌 목록 중에서 40.44%인 167건을 추출하였다.

표 4.8 세 번째 항목별가중치 세트 적용 유효 특허문헌 수

Table 4.8 The number of valid patent documents applied 3rd field weight set

국가	유사 특허문헌 수	수작업		세 번째 세트 항목별가중치	
		유효특헌	비율	유효특헌	비율
한국	754	266	35.28%	288	38.20%
일본	939	355	37.81%	345	36.74%
미국	1580	583	36.90%	624	39.49%
유럽	620	257	41.45%	261	42.10%
PCT	413	152	36.80%	167	40.44%
합계	4306	1613	37.46%	1685	39.13%

전체적으로는 유사 특허문헌 4306건 중에서 유효 특허문헌을 결정한 결과가 1685건의 유효 특허문헌을 추출하여 39.13%의 비율로 수작업을 이용한 경우보다 72건이 많다. 다만 일본의 경우만 10건 적게 추출되었다.

표 4.9는 특허조사원이 유효특허를 수작업으로 추출한 결과와 네 번째 항목별가중치 세트를 추출시스템에 적용하여 유효 특허문헌을 결정한 결과이다. 한국은 유사 특허문헌 목록 754건 중에서 258건으로 34.22%에 해당하며, 일본은 939건의 유사 특허문헌에서 34.61%의 비율인 325건이 추출되었고, 미국은 1580건의 유사 특허문헌 목록에서 591건으로 37.41%의 비율을 나타나며, 유럽은 유사 특허문헌 620건 중에서 233건으로 37.58%이고, PCT는 유효 특허문헌 목록 413건 중에서 156건으로 37.77% 비율이다.

전체는 4306건의 유사 특허문헌 목록에서 유효 특허문헌을 결정한 결과가 1563건의 유효 특허문헌을 추출하여 36.30%의 비율로 수작업을 이용한 경우보다 50건이 적다. 미국과 PCT는 수작업으로 유효 특허문헌을 추출한 건보다 많다.

표 4.9 네 번째 항목별가중치 세트 적용 유효 특허문헌 수

Table 4.9 The number of valid patent documents applied 4th field weight set

국가	유사 특허문헌 수	수작업		네 번째 세트 항목별가중치	
		유효특헌	비율	유효특헌	비율
한국	754	266	35.28%	258	34.22%
일본	939	355	37.81%	325	34.61%
미국	1580	583	36.90%	591	37.41%
유럽	620	257	41.45%	233	37.58%
PCT	413	152	36.80%	156	37.77%
합계	4306	1613	37.46%	1563	36.30%

표 4.10은 특허조사원이 수작업을 통해 유효특허를 추출한 결과와 다섯 번째 항목별 가중치 세트를 추출시스템에 적용하여 유효 특허문헌을 결정한 결과를 나타낸다. 국가별로는 한국의 경우, 754건의 유사 특허문헌 목록에서 251건으로 33.29%에 해당하며, 일본은 939건의 유사 특허문헌에서 309건으로 32.91%의 비율이고, 미국은 1580건의 유사 특허문헌 목록에서 630건으로 39.87%의 비율이며, 유럽은 유사 특허문헌 620건 중에서 196건으로 31.61%이고, PCT는 413건의 유효 특허문헌 목록에서 145건을 추출하여 35.11% 비율이다.

전체 4306건의 유사 특허문헌 중에서 유효 특허문헌을 결정한 결과가 1531건의 유효 특허문헌을 추출하여 35.56%의 비율로 수작업을 이용한 경우보다 82건이 적다. 하지만 미국의 경우에는 추출시스템을 통한 유효 특허문헌의 건수가 많다.

표 4.10 다섯 번째 항목별가중치 세트 적용 유효 특허문헌 수

Table 4.10 The number of valid patent documents applied 5th field weight set

국가	유사 특허문헌 수	수작업		다섯 번째 세트 항목별가중치	
		유효특헌	비율	유효특헌	비율
한국	754	266	35.28%	251	33.29%
일본	939	355	37.81%	309	32.91%
미국	1580	583	36.90%	630	39.87%
유럽	620	257	41.45%	196	31.61%
PCT	413	152	36.80%	145	35.11%
합계	4306	1613	37.46%	1531	35.56%

마지막으로 표 4.11은 특허조사원이 수작업을 통해 유효특허를 추출한 결과와 여섯 번째 항목별가중치 세트를 추출시스템에 적용하여 유효 특허문헌을 결정한 결과를 나타낸다. 국가별로는 한국의 경우, 754건의 유사 특허문헌 목록에서 267건으로 35.41%에 해당하며, 일본은 939건의 유사 특허문헌에서 371건으로 39.51%의 비율이고, 미국은 1580건의 유사 특허문헌 목록에서 644건으로 40.76%의 비율이며, 유럽은 유사 특허문헌 620건 중에서 248건으로 40.00%이다. 마지막으로 PCT는 413건의 유효 특허문헌 목록에서 148건을 추출하여 35.84% 비율이다.

표 4.11 여섯 번째 항목별가중치 세트 적용 유효 특허문헌 수

Table 4.11 The number of valid patent documents applied 6th field weight set

국가	유사 특허문헌 수	수작업		여섯 번째 세트 항목별가중치	
		유효특헌	비율	유효특헌	비율
한국	754	266	35.28%	267	35.41%
일본	939	355	37.81%	371	39.51%
미국	1580	583	36.90%	644	40.76%
유럽	620	257	41.45%	248	40.00%
PCT	413	152	36.80%	148	35.84%
합계	4306	1613	37.46%	1678	38.97%

전체적으로는 4306건의 유사 특허문헌 목록에서 유효 특허문헌을 결정한 결과가 1678건의 유효 특허문헌을 추출하여 38.97%의 비율로 수작업을 이용한 경우보다 65건이 많다. 일반적으로 특허동향분석에서 유사 특허문헌 목록으로부터 유효 특허문헌의 비율은 약 20%내외이다. 하지만 본 실험의 대상이 되는 방사성의약품 이용기술 개발 분야는 평균이 약 36.95%로 비교적 높다. 이는 본 실험대상 기술이 방사선기술을 이용한 치료

에 관한 것으로, 연구개발기술 분야에 연구하는 연구원이나 기업의 범위가 한정되고, 특허 출원이 많은 편이 아니어서 유사 특허문헌 목록에 비해 유효 특허문헌의 비율이 높다.

핵심 특허문헌 추출시스템의 정확률을 판단하기 위하여 특허조사원이 수작업을 통하여 추출한 유효 특허문헌을 대상으로 추출시스템을 통해 추출한 유효 특허문헌과 비교하여 얼마나 동일한지 여부를 비교 평가한다. 국가별 및 전체의 정확률을 함께 확인하여 항목별가중치 세트에 따라 전체의 정확률뿐만 아니라 국가별 정확률의 변화도 분석하고자 한다. 먼저 첫 번째 항목별가중치 세트를 적용하여 추출된 유효 특허문헌 중에서 수작업을 통한 유효 특허문헌과의 일치되는 문헌의 수 및 정확률이 그림 4.13과 같다.

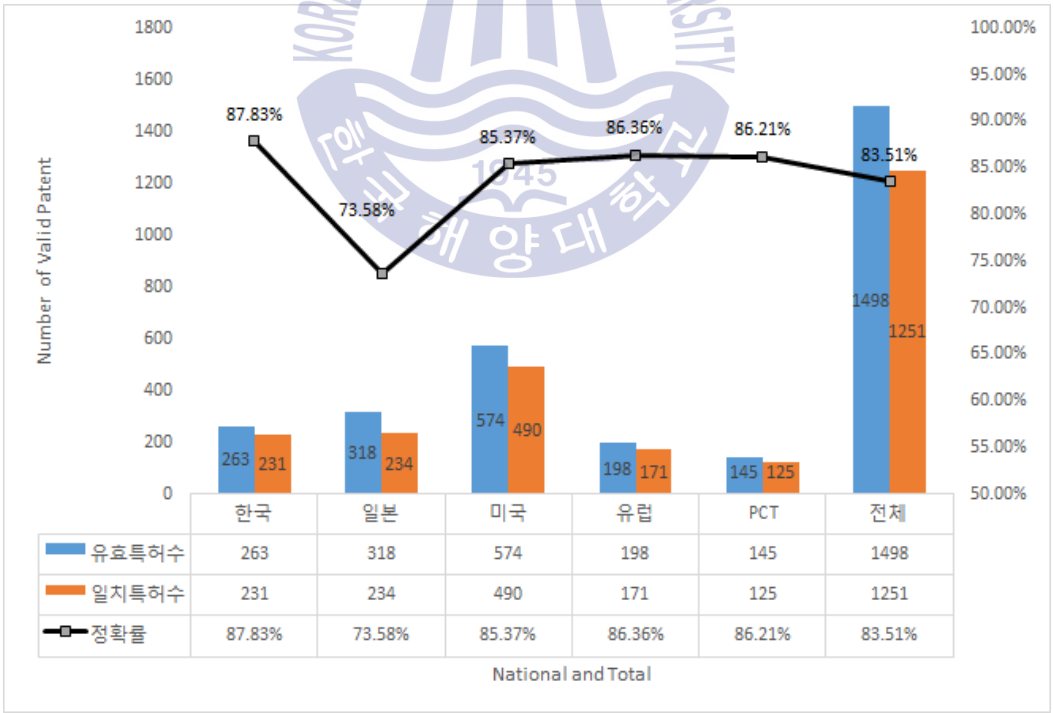


그림 4.13 첫 번째 항목별가중치 세트 적용 유효 특허문헌의 정확률

Fig. 4.13 The accuracy of valid patent documents applied 1st field weight set

한국의 경우에는 유사도 측정에 의해 263건이 유효 특허문헌으로 추출되었으나, 231건이 수작업을 통한 유효 특허문헌과 일치하고, 32건이 불일치하여 87.83%의 정확률을 나타낸다. 일본은 318건의 추출된 유효 특허문헌 중에서 234건이 일치하고, 84건이 불일치하여 73.58%의 정확률이 보인다.

미국의 경우에는 574건의 유효 특허문헌 중에서 490건의 유효 특허문헌이 일치하고, 84건이 불일치하여 85.37%가 정확하며, 유럽의 경우에는 유효 특허문헌 198건 중 171건이 일치하고 27건이 불일치하여 86.36%의 정확률이 보이고, PCT는 145건의 추출된 유효 특허문헌에서 125건이 일치한 반면 20건이 불일치하여 86.21%의 정확률을 나타낸다.

전체적으로는 유효 특허문헌 1498건 중에서 1251건이 일치하고, 247건이 불일치하여 83.51%의 정확률이 보이고 있다. 상대적으로 한국의 경우에는 높은 정확률을 나타내고 일본은 가장 낮은 정확률이 나타남으로서 기술요약서에서 추출한 색인후보어가 기술내용을 충실하게 표현하는 단어였음을 알 수 있다. 이에 반해 일본 특허문헌은 한국어로 기계번역한 결과에 따른 표현 단어의 부정확함을 보여주고 있다.

다음으로 그림 4.14는 두 번째 항목별가중치 세트가 적용된 유효 특허문헌의 국가별 및 전체 유효특허 문헌의 수와 일치하는 문헌의 수 및 정확률을 나타낸다. 한국은 유사도 측정에 의해 295건이 유효 특허문헌으로 추출되었으나, 256건이 수작업을 통한 유효 특허문헌과 일치하고, 39건이 불일치하여 86.78%의 정확률을 나타낸다.

일본은 346건의 추출된 유효 특허문헌 중에서 247건이 일치하고, 99건이 불일치하여 71.39%의 정확률이 보인다. 미국의 경우에는 633건의 유효 특허문헌 중에서 543건의 유효 특허문헌이 일치하고, 99건이 불일치하여 85.78%가 정확하다. 유럽의 경우에는 유효 특허문헌 255건 중 216건이 일치하고 39건이 불일치하여 84.71%의 정확률이 보이고, PCT는 162건의 추출된 유효 특허문헌에서 138건이 일치한 반면 24건이 불일치하여 85.19%의 정확률을 나타낸다.

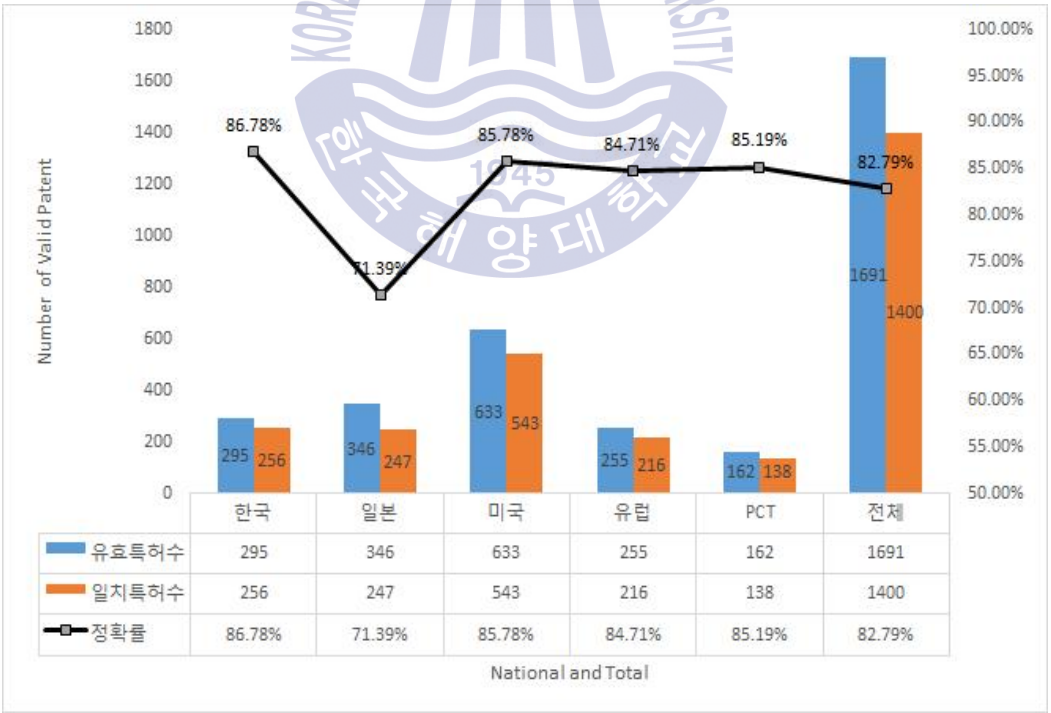


그림 4.14 두 번째 항목별가중치 세트 적용 유효 특허문헌의 정확률

Fig. 4.14 The accuracy of valid patent documents applied 2nd field weight set

전체 추출된 유효 특허문헌 1691건 중에서 1400건이 일치하고, 291건이 불일치하여 82.79%의 정확률이 보이고 있다. 첫 번째 항목별가중치 세트와 비슷한 경향을 보이고 있으나 일본과 유럽이 상대적으로 더욱 정확률이 낮아져 전체 정확률이 더 낮은 것으로 나타나는 경향을 보이고 있다.

세 번째 항목별가중치 세트를 적용하여 추출된 유효 특허문헌 중에서 수작업을 통한 유효 특허문헌과의 일치되는 문헌의 수를 국가별 및 전체 수와 정확률을 비교한 결과는 그림 4.15에 도시된 바와 같다. 한국의 경우에는 추출된 288건이 유효 특허문헌 중에서 253건이 수작업을 통한 유효 특허문헌과 일치하고, 35건이 불일치하여 87.85%의 정확률을 나타낸다. 일본의 경우, 345건의 추출된 유효 특허문헌 중에서 251건이 일치하고, 94건이 불일치하여 72.75%의 정확률이 보인다.

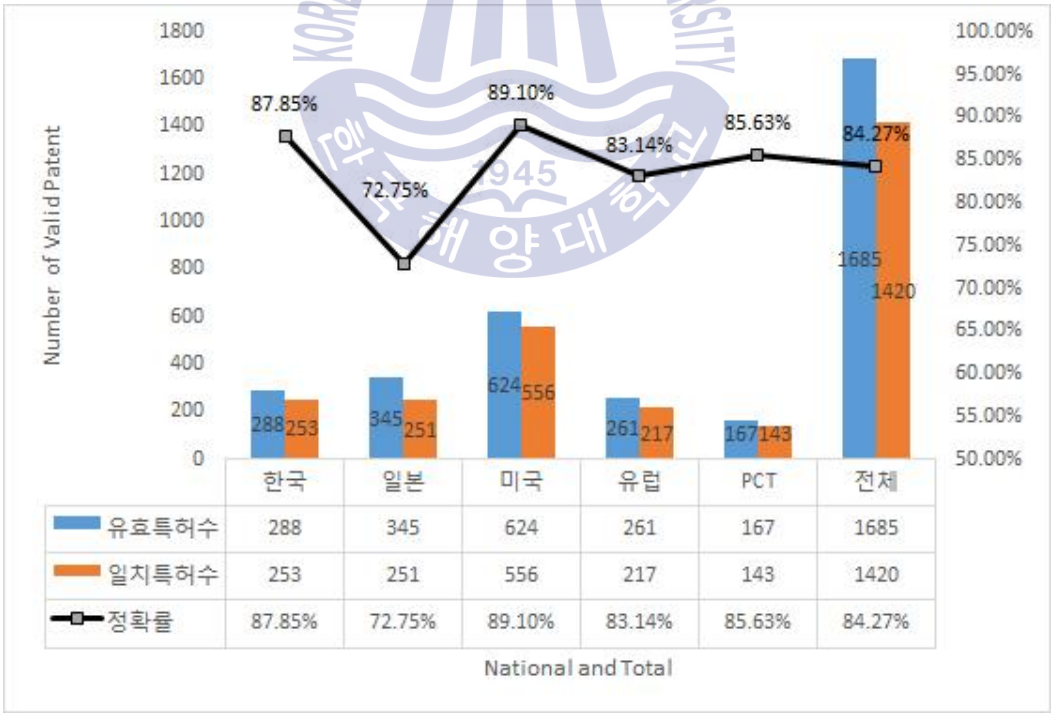


그림 4.15 세 번째 항목별가중치 세트 적용 유효 특허문헌의 정확률

Fig. 4.15 The accuracy of valid patent documents applied 3rd field weight set

미국의 경우에는 624건의 유효 특허문헌 중에서 556건의 유효 특허문헌이 일치하고, 68건이 불일치하여 89.10%가 정확하며, 유럽은 유효 특허문헌 261건 중 217건이 일치하고 44건이 불일치하여 83.14%의 정확률이 보이고, PCT의 경우에는 167건의 추출된 유효 특허문헌에서 143건이 일치한 반면 24건이 불일치하여 85.63%의 정확률을 나타낸다.

전체 유효 특허문헌 1685건 중에서 1420건이 일치하고, 265건이 불일치하여 84.27%의 정확률이 보이고 있다. 첫 번째 세트 및 두 번째 항목별가중치 세트에 비해 세 번째 항목별가중치 세트는 비교적 미국의 정확률이 가장 높은 경향을 보였으며 상대적으로 유럽의 정확률이 낮아진 특징이 있고, 그 외는 비슷한 경향을 보이고 있다.

그림 4.16은 네 번째 항목별가중치를 적용하여 추출된 유효 특허문헌 중에서 수작업을 통한 유효 특허문헌과의 일치되는 문헌의 수와 정확률을 국가별 및 전체 수를 비교한 결과를 나타낸다. 한국의 경우에는 유사도 측정에 의해 258건이 유효 특허문헌으로 추출되었으나, 232건이 수작업을 통한 유효 특허문헌과 일치하고, 26건이 불일치하여 89.92%의 정확률을 나타낸다.

일본은 325건의 추출된 유효 특허문헌 중에서 235건이 일치하고, 90건이 불일치하여 72.30%의 정확률이 보인다. 미국의 경우에는 591건의 유효 특허문헌 중에서 536건의 유효 특허문헌이 일치하고, 55건이 불일치하여 90.69%가 정확률을 나타낸다.

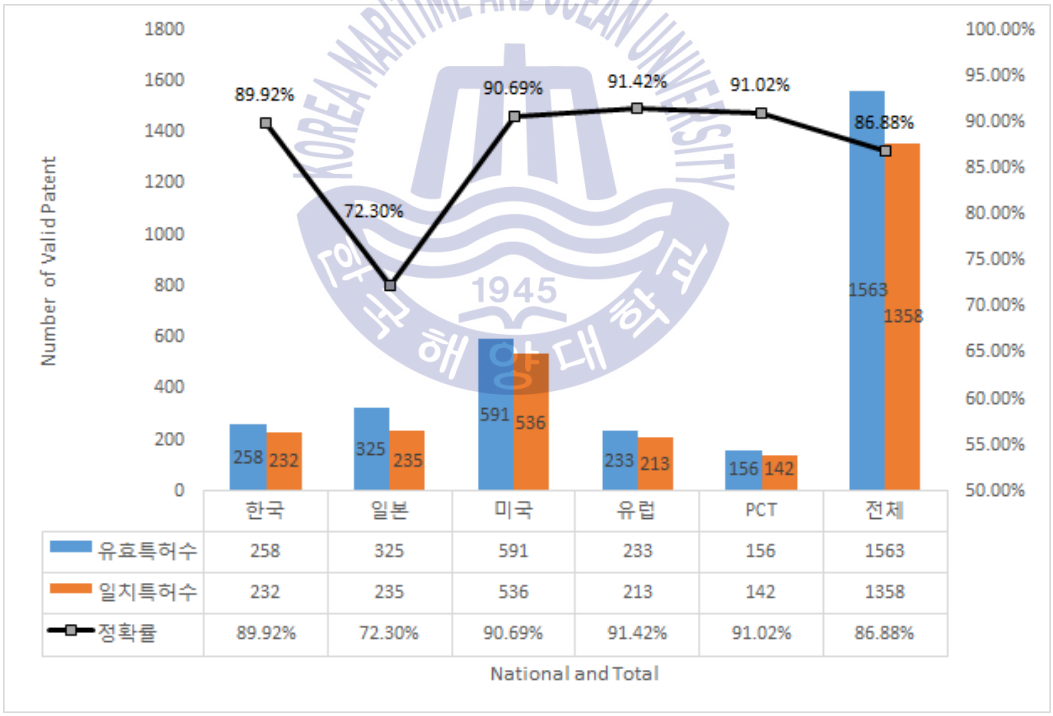


그림 4.16 네 번째 항목별가중치 세트 적용 유효 특허문헌의 정확률

Fig. 4.16 The accuracy of valid patent documents applied 4th field weight set

유럽의 경우에는 유효 특허문헌 233건 중 213건이 일치하고 20건이 불일치하여 91.42%의 정확률이 보이고, PCT는 156건의 추출된 유효 특허문헌에서 142건이 일치한 반면 14건이 불일치하여 91.02%의 정확률을 나타낸다.

전체적으로는 유효 특허문헌 1563건 중에서 1358건이 일치하고, 205건이 불일치하여 86.88%의 정확률이 보이고 있다. 네 번째 항목별가중치 세트는 일본의 경우를 제외하면 이전 세 개의 항목별가중치 세트에 비해 향상된 결과를 보인다. 특히 영어검색식 기반의 미국, 유럽, PCT 특허문헌이 높은 정확률을 보이고 있다. 이는 패밀리수가 많은 일부 특허문헌의 유사도가 높아진 영향이 반영된 결과이다.

그림 4.17은 다섯 번째 항목별가중치 세트가 적용된 유효 특허문헌의 국가별 및 전체 유효특허 문헌의 수와 일치하는 문헌의 수 및 정확률을 나타낸다. 한국은 유사도 측정에 의해 251건이 유효 특허문헌으로 추출되었으며, 225건이 수작업을 통한 유효 특허문헌과 일치하고, 26건이 불일치하여 89.64%의 정확률을 나타낸다. 일본의 경우에는 309건의 추출된 유효 특허문헌 중에서 222건이 일치하고, 87건이 불일치하여 71.84%의 정확률이 보인다.

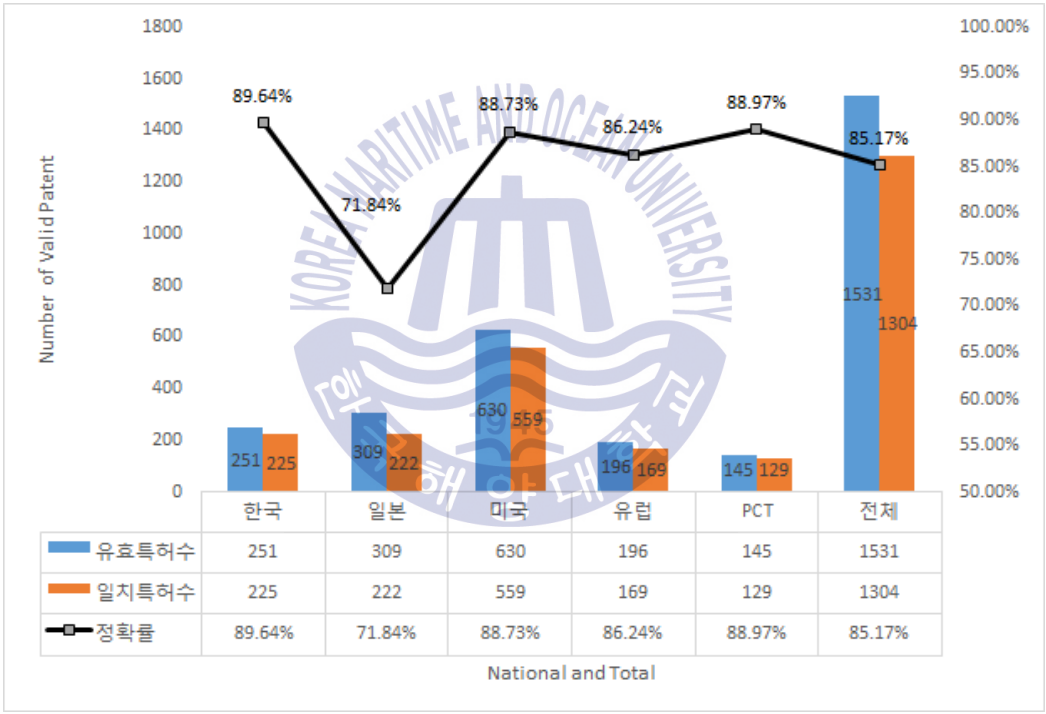


그림 4.17 다섯 번째 항목별가중치 세트 적용 유효 특허문헌의 정확률

Fig. 4.17 The accuracy of valid patent documents applied 5th field weight set

미국의 경우, 630건의 유효 특허문헌 중에서 559건의 유효 특허문헌이 일치하고, 71건이 불일치하여 88.73%가 정확하며, 유럽의 경우에는 유효 특허문헌 196건 중 169건이 일치하고 27건이 불일치하여 86.24%의 정확률이 보인다. 그리고 PCT는 145건의 추출된 유효 특허문헌에서 129건이 일치한 반면 16건이 불일치하여 88.97%의 정확률을 나타낸다.

전체적으로는 유효 특허문헌 1531건 중에서 1304건이 일치하고, 227건이 불일치하여 85.17%의 정확률이 보이고 있다. 다섯 번째 항목별가중치 세트는 한국은 높은 정확률을 유지하였으나 유럽의 특허문헌의 정확률이 낮은 경향을 보이고 있다.

마지막으로 그림 4.18은 여섯 번째 항목별가중치 세트를 적용하여 추출된 유효 특허문헌 중에서 수작업을 통한 유효 특허문헌과의 일치되는 문헌의 수 및 정확률을 국가별 및 전체로 구분하여 나타낸다.

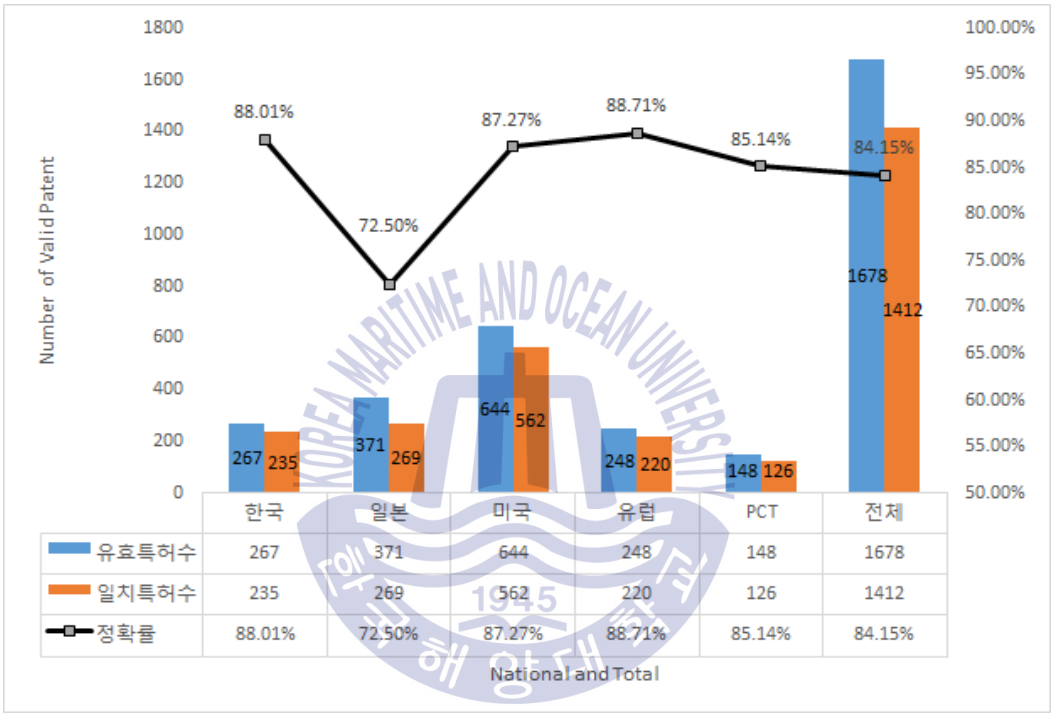


그림 4.18 여섯 번째 항목별가중치 세트 적용 유효 특허문헌의 정확률

Fig. 4.18 The accuracy of valid patent documents applied 6th field weight set

한국의 경우, 267건의 추출된 유효 특허문헌 중에서 235건이 수작업을 통한 유효 특허문헌과 일치하고, 32건이 불일치하여 88.01%의 정확률을 나타낸다. 일본의 경우, 371건의 추출된 유효 특허문헌 중에서 269건이 일치하고, 102건이 불일치하여 72.50%의 정확률이 보인다. 미국은 644건의 유효 특허문헌 중에서 562건의 유효 특허문헌이 일치하고, 82건이 불일치하여 87.27%가 정확률을 나타낸다.

유럽의 경우에는 유효 특허문헌 248건 중 220건이 일치하고 28건이 불일치하여 88.71%의 정확률이 보이고, PCT는 148건의 추출된 유효 특허문헌에서 126건이 일치한 반면 22건이 불일치하여 85.14%의 정확률을 나타낸다. 전체적으로는 유효 특허문헌 1678건 중에서 1412건이 일치하고, 288이 불일치하여 84.15%의 정확률이 보이고 있다. 여섯 번째 항목별가중치 세트는 다른 항목별가중치에 비해 유럽과 일본이 높은 정확률을 나타내고 있는 경향을 보이고 있다.

일반적으로 특허동향분석에서 후보 특허문헌 대비 유효 특허문헌의 비율은 약 20%내외이다. 하지만 방사성의약품 이용기술 개발 분야는 4306건의 후보 특허문헌을 대상으로 1498건에서 1691건으로 평균 36.95%로 비교적 높다. 이는 본 실험대상인 방사선의약품 분야는 방사선기술을 이용한 치료 및 진단에 관한 것으로, 해당 분야에 연구하는 연구원이나 기업의 범위가 한정되고, 특허 출원이 많은 편이 아니어서 후보 특허문헌의 수에 비해 유효 특허문헌 수의 비율이 높다.

특허조사원의 수작업을 통하여 추출한 유효 특허문헌을 기준으로 추출시스템을 통해 추출된 유효 특허문헌에 대하여 항목별가중치 각 세트별 정확률은 그림 4.19와 같이 정리된다. 전체적으로 약 83% 이상의 정확률을 나타내고 있으며, 특히 넷 세트의 항목별 가중치가 가장 높은 정확률인 86.88%를 보이고 있다.

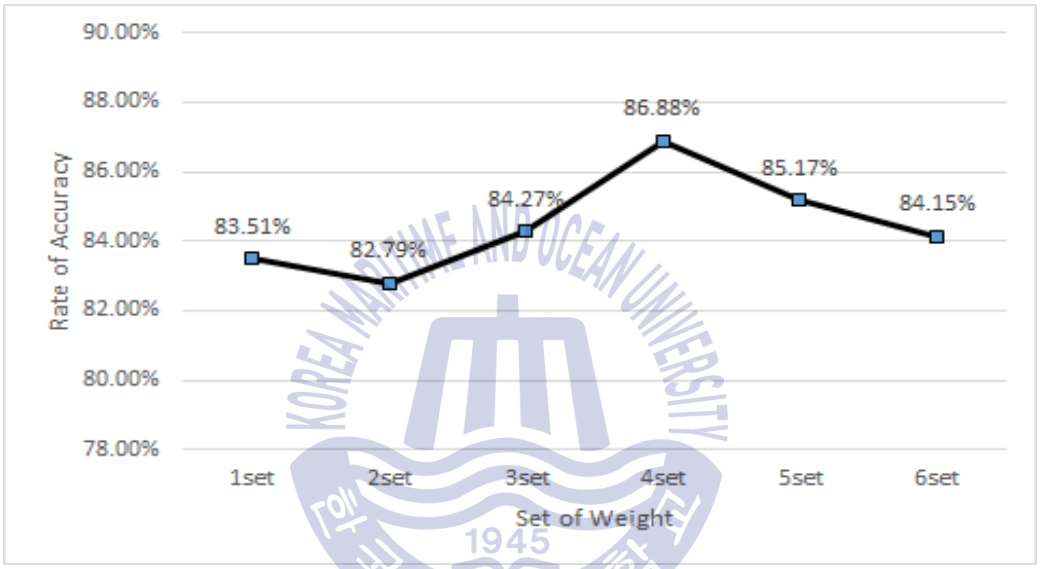


그림 4.19 항목별가중치 세트별 유효 특허문헌의 정확률

Fig. 4.19 The accuracies of valid patent documents by sets of field weight

특허조사원이 수작업을 통해 검색된 후보 특허문헌의 목록에서 유효 특허문헌을 추출하는데 필요한 시간과 추출시스템을 이용하여 유효 특허문헌을 추출하는데 소요되는 시간을 비교하여 성능을 평가한다. 본 논문에서는 특허검색서비스로부터 추출한 4306건의 후보 특허문헌의 목록을 대상으로 실험하였다.

표 4.12는 방사성의약품 이용기술 개발 분야에 대한 유효 특허문헌을 추출하는데 수작업을 통하여 추출하는데 필요한 수행시간과 추출시스템을 이용하여 추출한 수행시간을 비교하였다. 일반적으로 특허조사원이 영어로 기재되어 있는 미국, 유럽 및 PCT 특허문헌을 파악하는 시간은 한글로 기재되어 있는 한국과 일본의 특허문헌을 파악하는 시간보다 3배 이상의 시간이 소요된다. 그리고 미국이나 유럽이 선진기술인 분야의 경우에는 핵심 특허문헌의 기초가 되는 유효 특허문헌의 추출에 더욱 많은 시간이 요구된다. 또한 일본 특허문헌의 경우에는 기계번역을 통한 한글 문헌이기 때문에 기재된 표현의 부정확으로 인해 기술내용을 파악하는데 어려움이 있다.

표 4.12 유효 특허문헌 추출 수행시간 비교

Table 4.12 The comparison of performance time for extracting valid patent documents

국가	후보 특허문헌 수	수작업		추출시스템 (네 번째 세트 기준)	
		유효특헌	시간	유효특헌	시간
한국	754	266	732분	258	6분
일본	939	355	1,342분	325	7분
미국	1,580	583	5,000분	591	13분
유럽	620	257	1,962분	233	6분
PCT	413	152	1,307분	156	4분
합계	4,306	1,613	10,343분	1563	36분

본 실험의 대상이 되는 방사선의약품 분야도 화학 및 생물 분야로 미국이나 유럽이 기술의 선진국으로 기술내용을 확인하는데 많은 시간이 소요된다. 그리고 검색된 후보 특허문헌의 건수도 미국이 가장 많고, 유럽도 작은 편이 아니며, 일본 특허문헌도 많이 검색되어 수작업을 경우 시간이 많이 걸린다. 다만 해당 기술 분야가 특허조사원의 전공분야인 경우에는 시간을 많이 축소될 수 있다.

이에 반해 추출시스템을 이용한 유효 특허문헌의 추출은 기술요약서와 특허문헌의 문서벡터를 이용한 유사도 측정을 이용하므로 특허문헌의 건수와 추출시스템의 성능에 따른 수행시간의 차이만 있을 뿐이다. 그리고 유사도 측정은 복잡한 연산을 수행하는 것이 아니기 때문에 많은 시간이 소요되지 않는다. 다만 연산된 결과로부터 유사도 값을 기반으로 순위화하여 추출하는데 약간의 시간이 필요하다.

따라서 특허조사원이 수작업으로 전체 4036건에서 1613건의 유효 특허문헌을 추출하는데 10,343분 즉 약 172.38 시간이 필요하였으나, 추출시스템으로 1563건의 유효 특허문헌을 추출하는 데는 약 36분 정도의 시간이 소요되어 약 287배의 차이가 발생한다.

4.5.2 세부기술별 특허문헌 분류

방사성의약품 이용기술은 방사성의약품 개발, 진단 및 치료 성능 평가 기술, 질환모델 생산 기술로 분류된다. 기술분류의 정확성을 확인하기 위하여 유사도측정을 통한 1563건의 유효 특허문헌이 아닌 수작업을 통해 추출된 1613건의 유효특허문헌을 대상으로 신경망 알고리즘을 이용하여 기술분류를 수행한다.

첫 번째 기술분야인 방사성의약품 개발은 난치성 질환인 암, 뇌질환 등에 대한 진단 및 치료를 위한 방사성 의약품이나 방사성 동위원소와 관련된 기술이다. 두 번째 기술분야인 진단 및 치료 성능 평가 기술은 난치성 질환의 진단이나 치료 성능을 파악하기 위한 영상 장비 및 진단 약물의 변화량 측정과 관련된 기술이다. 마지막으로, 질환 모델 생산 기술은 동물을 대상으로 약물의 시험에 관한 기술이다.

핵심 특허문헌 추출시스템의 기술분류를 위한 신경망 입력층의 입력변수값 xi는 기술요약서에서 추출된 131개의 색인후보어의 각 유효 특허문헌에 대한 TF-IDF가중치이다. 기술분류와 관련된 색인후보어 목록은 표 4.13과 같다. 따라서 표 4.13에 기재되어 있는 색인후보어에 대한 분류가중치 v_i는 2.0이고, 그 외 색인후보어는 1.0이다.

표 4.13 기술 관련 색인후보어들

Table 4.13 The index candidates to technology-related

기술분류	색인후보어
방사성의약품 개발	방사성의약품, 방사성동위원소, 치료, 항체, 나노입자, 추적, 줄기세포, 면역세포, 테라그노시스, 신약, 의약품개발기술, 표적치료, 나노입자, 바이오시밀러, 면역세포, 줄기세포, 의약품, 바이오신약, 마이크로도징
진단 및 치료 성능 평가 기술	영상, 최적화, 분석, 분자영상, 영상기반, 정량, 이중타겟, 생체분포, 생물학적, 모니터링, 양전자방출단층촬영, 방사성표지, 화학적, 흡수선량평가, 영상보정, 정량분석기술, 표적치료제, 유효선량, 영상진단, 등가균일선량, 대사종양부피
질환모델 생산 기술	종양, 질환, 난치성질환, 모델, 뇌질환, 질환모델, 암질환, 동물모델

방사성의약품 이용기술 개발의 세부기술별로 유효 특허문헌을 분류함에 있어, 학습데이터와 검증데이터는 표 4.14에서와 같이, 원자력관련 연구소의 연구원이 유효 특허문헌 중에서 각 분야별로 직접 분류하였던 데이터를 학습데이터로 이용하였다. 원자력연구원이 제시한 방사성의약품 개발과 관련한 98건, 진단 및 치료 성능 평가 기술과 관련한 107건, 질환모델 생산 기술 분야와 관련한 51건으로 총 256건의 학습데이터로 이용하였다. 그리고 그 외 방사성의약품 개발과 관련한 424건, 진단 및 치료 성능 평가 기술과 관련한 602건, 질환모델 생산 기술 분야와 관련한 331건으로 총 1357건을 검증데이터로 실험하였다.

표 4.14 기술분류용 학습데이터 및 검증데이터의 수

Table 4.14 The number of training data and test data for technical classification

기술분류	학습데이터 (비율)	검증데이터 (비율)	합계
방사성의약품 개발	98 (18.8%)	424 (81.2%)	522
진단 및 치료 성능 평가 기술	107 (15.1%)	602 (84.9%)	709
질환모델 생산 기술	51 (13.4%)	331 (86.6%)	382
합계	256	1357	1613

학습데이터가 풍부하지 않은 관계로 학습데이터를 통해서 학습과정을 거친 후 검증데이터를 분류함에 있어 분류된 값 중에서 가장 높은 상위 25%씩을 다시 학습데이터로 이용하면서 표 4.15와 같이, 4회에 거쳐 추출시스템을 이용하여 실험하였다. 제안한 신경망 구조에서 각 특허문헌마다 색인후보어 131개의 TF-IDF가중치를 입력으로 실험하였으며, 각 회차별로 기술분류를 수행한 검증데이터 중에서 상위 25%인 약 339개를 다시 학습데이터로 이용하므로 2회 및 3회에 학습데이터 추가 및 시스템 수정에 추가적인 시간이 필요하였다. 마지막 4회에서는 검증데이터 중 학습데이터로 이용되는 데이터가 없으므로 빠르게 수행하였다.

표 4.15 추출시스템의 기술분류 수행시간

Table 4.15 The performance time for technical classification of extracting system

	1회	2회	3회	4회
학습데이터	256	595	934	1273
검증데이터	1357	1018	679	340
상위 25%의 수	339	339	339	0
수행시간	12분	25분	23분	13분

특허조사원이 1613건의 유효 특허문헌에 대하여 국가별로 수작업을 통하여 세부기술별로 분류를 수행하는데 필요한 시간이 표 4.16과 같다. 동일 또는 유사한 특허문헌인 패밀리특허가 비교적 많았던 미국, 유럽 특허문헌의 경우에는 영어 문헌임에도 시간을 줄일 수 있었다. 하지만 추출시스템의 경우에는 세부기술별 분류 절차를 처리하는데 약 73분으로 2,243분이 걸린 수작업보다 약 31배가 빠른 성능을 나타낸다.

표 4.16 특허조사원의 기술분류 수행시간

Table 4.16 The performance time for technical classification of patent researchers

국가	수작업	
	유효특헌	시간
한국	266	163분
일본	355	252분
미국	583	1066분
유럽	257	485분
PCT	152	277분
합계	1,613	2,243분

방사성의약품 이용기술 개발과 관련하여, 전체 데이터에 대하여 원자력관련 연구소의 연구원이 확인한 실제 기술분류와 본 핵심 특허문헌 추출시스템의 신경망 알고리즘을 이용한 기술분류의 결과는 표 4.17과 같다. 신경망 알고리즘을 이용한 기술분류의 전체 정확률은 91.08%로 확인되었다.

그러나 기술분류별로 정확률의 상위는 질환모델 생산 기술, 방사성의약품 개발이며, 진단 및 치료 성능 평가 기술이 가장 낮다. 이는 기술분류 관련 색인후보어가 나타내는 기술이 보다 명확한 질환모델 생산기술은 높은 정확률을 보인 반면, 일반적인 영상장치 관련 색인후보어를 포함하는 진단 및 치료 성능 평가 기술은 낮은 정확률을 보였다.

표 4.17 추출시스템의 기술분류 정확률

Table 4.17 The accuracy of technical classification of the extraction system

기술분류	수작업	학습데이터	검증데이터		정확률
			일치	불일치	
방사성의약품 개발	522	98	386	38	91.04%
진단 및 치료 성능 평가 기술	709	107	541	61	89.87%
질환모델 생산 기술	382	51	309	22	93.35%
합계	1613	256	1236	121	91.08%

이는 특허 문서 분류 알고리즘 비교 연구[71, 74]에 기재된 다른 알고리즘의 정확률과 비교한 결과는 표 4.18에서와 같다. 여기서 그 대상이 IPC코드의 경우보다 특정 기술분야를 대상으로 하는 경우에 비교적 향상된 결과가 나왔다. 그리고 특정 기술분야를 대상으로 한 경우에도 본 추출시스템의 경우 분류 관련된 색인후보어에 대한 분류가중치를 직접적으로 부가하여 좋은 결과를 얻을 수 있었다.

표 4.18 기술분류 알고리즘간 정확률 비교

Table 4.18 The comparison of accuracies among technical classification algorithms

알고리즘	대상	정확률
SVM	IPC코드	88.90%
용어클러스터링	IPC코드	84.97%
*k*NN	Wearable IoT	90.83%
나이브 베이지안	Wearable IoT	93.12%
Linear SVM	Wearable IoT	90.37%
제안한 시스템	방사성의약품	91.08%

4.5.3 핵심 특허문헌 추출

핵심 특허는 방사성의약품 이용기술의 하위 기술분류에 속하는 유효 특허문헌 중에서 각 기술분야별로 연구개발 기술과 유사도가 높고, 기술에 있어 원천기술 또는 핵심기술이며, 지식재산권의 관점에서 권리가 높은 특허문헌에 해당한다. 따라서 향후 연구개발 기술과의 관계에서 그 내용을 보다 상세히 정성분석을 하기 위해서 추출되는 특허문헌이다.

방사성의약품 이용기술의 세부기술별로 분류된 각 유효 특허문헌로부터 표 3.5와 표 3.6의 핵심 특허문헌 선정기준에 해당하는 항목값과 우선순위에 따른 세트별 우선순위 가중치를 이용하여 식 (3.20)을 이용하여 선정값을 연산한다.

그림 4.20과 같이, 특허번호를 유효 특허문헌의 구분자로 핵심 특허문헌 추출을 위한 유사도 등의 각 항목과 해당 값을 나타내고, 각 세부기술별 분류된 유효 특허문헌의 항목에 있어 선정값 항목을 추가하여 연산된 선정값을 입력한다. 입력된 선정값을 기준으로 각 세부기술별로 유효 특허문헌을 순위화하여 핵심 특허문헌을 추출한다.

번호	유사도	유사도*10	특허평가점수	등록여부	패밀리국가	패밀리수	피인용수	독립청구항	권리만료여	선정값
JP2009501318T	0.568328	5.683283443	3	0	9	9	2	4	1	53.66657
KR20050039405A	0.555279	5.552790068	6	1	12	14	1	1	0	67.00558
JP2013521233T	0.504223	5.04222995	1	0	4	4	0	3	1	29.08446
KR20140062961A	0.455189	4.551892826	1	0	1	1	0	3	1	18.80379
KR20127016163A	0.378876	3.788756221	5	0	17	165	3	17	1	317.4775
KR20120142532A	0.337009	3.370092217	2	1	1	1	0	1	0	16.54018
KR20057010811A	0.335669	3.356690891	2	0	15	15	0	1	1	59.31338
KR20150061829A	0.319542	3.195417359	0	0	0	0	0	3	1	11.09083
KR20067012885A	0.312025	3.120254675	1	0	13	13	0	1	1	50.74051
JP2008540586T	0.241207	2.412072949	1	0	6	6	1	2	1	30.22415
KR20057010811A	0.234437	2.344368861	2	0	15	15	0	1	1	57.28874
KR20090063996A	0.210024	2.100243415	1	0	12	14	0	2	1	49.80049
KR20150181902A	0.197441	1.974407043	0	0	0	0	0	3	1	8.648814
JP2007512324T	0.192844	1.928435953	1	0	13	13	0	1	1	48.35687
KR20130110282A	0.176844	1.768444442	2	1	3	3	0	2	0	20.73689
KR20127016163A	0.146959	1.469593412	4	0	17	165	3	17	1	310.9392
KR20097001229A	0.146034	1.460336779	2	1	10	12	0	3	0	46.02067
KR20127033515A	0.114218	1.142175329	3	0	9	21	0	4	1	59.78435
JP2007537245T	0.074696	0.746961939	2	1	12	14	1	7	0	56.99392
KR20130074658A	0.073378	0.733783933	4	1	1	1	0	6	0	21.06757
JP2005534641T	0.070996	0.709958789	1	1	8	10	0	3	0	36.41992
KR20070096662A	0.062795	0.627945787	1	0	12	14	0	2	1	46.85589
KR20097001229A	0.039318	0.39318477	2	1	10	12	0	3	0	43.88637
JP2004113791A	0.032295	0.322947288	2	0	4	4	7	4	1	32.54589
JP2007512324T	0.027936	0.279357808	3	0	13	13	0	1	1	48.85872
JP2013521233T	0.027107	0.271070355	1	0	4	4	0	3	1	19.54214

그림 4.20 핵심 특허문헌의 추출 항목 및 선정값

Fig. 4.20 The patent features and selection values for core patent documents

핵심 특허문헌 추출을 위한 유효특허문헌은 수작업으로 세부기술별로 분류된 1613건의 유효특허문헌을 대상은 추출하였다. 수작업을 통한 핵심 특허문헌은 방사성의약품 개발 분야는 12건, 진단 및 치료 성능 평가 기술은 12건, 질환모델 생산 기술은 9건을 추출하였다. 추출시스템을 통하여 추출된 핵심 특허문헌의 수도 수작업을 통해 얻어진 핵심 특허문헌의 수와 같은 수로 추출하였다.

세부기술별로 핵심 특허문헌을 추출하기 위하여 특허조사원이 수작업을 통해 추출한 수행시간과 추출시스템을 통하여 각 항목별 값과 우선순위를 선형 연산한 후 정렬 후 추출하는데 필요한 수행시간은 표 4.19와 같은 시간이 요구되었다. 특허조사원이 수작업을 통해 핵심 특허문헌을 추출하는데 있어, 특허조사원이 유효 특허문헌 추출 및 세부기술별 분류를 통해 각 특허문헌에 대한 기술내용의 파악이 좀 더 잘 이루어져 있고, 유효 특허문헌의 추출에 비해 추출대상의 수가 적어 수작업의 경우에도 추출하는데 많은 시간이 요구되지는 않아 509분의 시간이 필요했다.

그러나 추출시스템을 이용하여 핵심 특허문헌 추출을 수행하는 경우에는 선정 기준이 명확하다. 그리고 선정기준에 해당하는 항목값과 우선순위가중치가 명확하게 정해져 있기 때문에 약 25분 정도의 시간으로 추출할 수 있어 수작업에 비해 약 21배 정도의 시간을 단축할 수 있었다.

표 4.19 핵심 특허문헌 추출 수행시간 비교

Table 4.19 The comparison of performance time for extracting core patent documents

기술분류	유효특허 문헌의수	핵심특허 문헌의 수	수작업 수행 시간	추출시스템 수행 시간
방사성의약품 개발	522	12	146분	8분
진단 및 치료 성능 평가 기술	709	12	201분	11분
질환모델 생산 기술	382	9	162분	6분
핵심특허수	1613	33	509분	25분

그리고 추출시스템을 통한 핵심 특허문헌 추출의 정확률을 판단하기 위하여 표 3.6에서 선정한 각 우선순위가중치 세트별로 실험을 수행하였다. 먼저 그림 4.21은 특허조사원이 수작업을 통해 추출한 핵심 특허문헌의 결과를 대상으로 첫 번째 우선순위가중치 세트를 적용한 핵심 특허문헌 중 일치하는 특허문헌과 오류특허문헌의 수 및 세부기술별 정확률을 보여준다.

첫 번째 우선순위가중치 세트를 적용하여 세부기술별로 핵심 특허문헌을 추출한 결과 방사선의약품 개발 분야는 전체 12건 중에서 9건의 특허문헌이 일치하였고, 3건의 특허문헌이 불일치하여 75%의 정확률을 보이고 있다. 진단 및 치료 성능 평가 기술 분야는 전체 12건 중에서 7건의 특허문헌이 일치하였고, 5건의 특허문헌이 불일치하여 정확률이 58.33%이다.

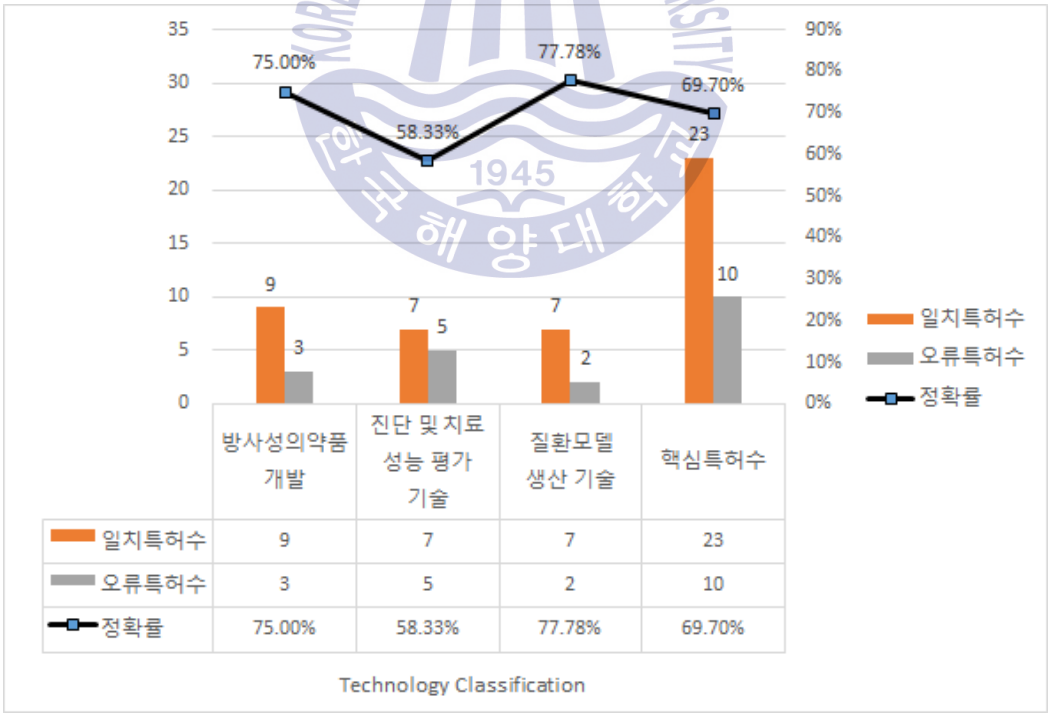


그림 4.21 첫 번째 우선순위가중치 세트 적용 핵심 특허문헌의 정확률

Fig. 4.21 The accuracy of core patent documents applied 1st priority weight set

질환모델 생산기술 분야는 9건 중에서 7건의 특허문헌이 일치하였고, 2건의 특허문헌이 불일치하여 77.78%의 정확률을 나타내고 있다. 그리고 전체 33건의 핵심 특허문헌 중에서 23건의 특허문헌이 일치하고, 10건의 특허문헌이 불일치하여 정확률은 69.70%이다.

그리고 그림 4.22는 특허조사원이 수작업을 통해 추출한 핵심 특허문헌의 결과를 기준으로 두 번째 우선순위가중치 세트를 적용한 결과인 핵심 특허문헌 중 일치하는 특허문헌과 오류특허문헌의 수 및 핵심 특허문헌 추출의 세부기술별 정확률을 보여준다. 방사선의약품 개발 분야는 전체 12건 중에서 9건의 특허문헌이 일치하였고, 3건의 특허문헌이 불일치하여 75%의 정확률을 보이고 있다.

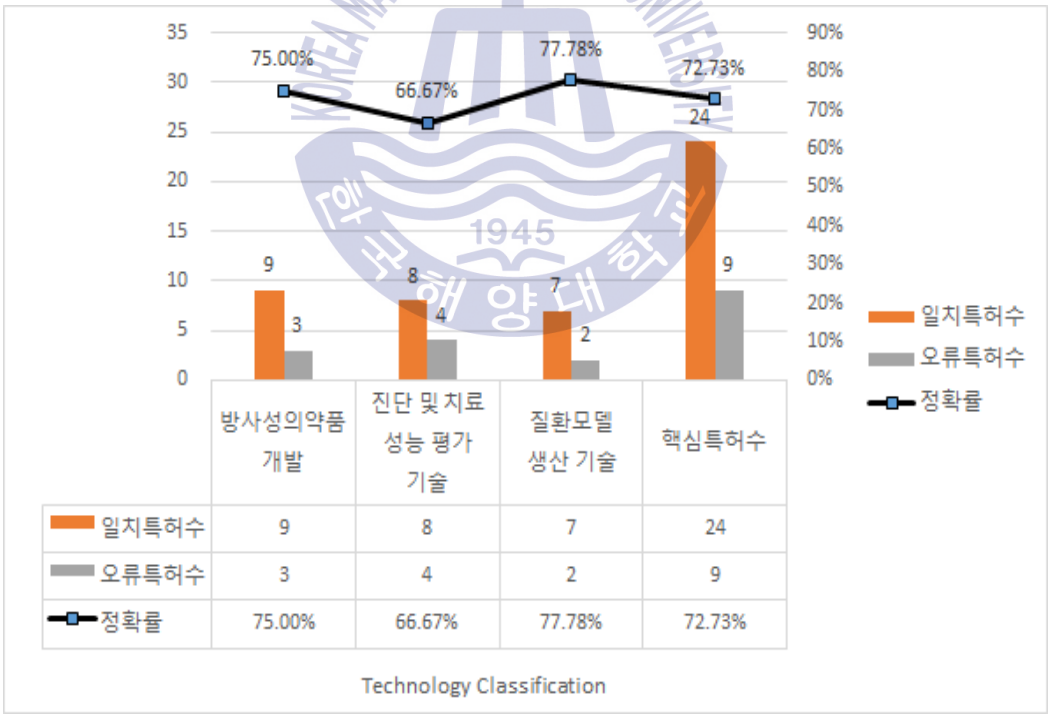
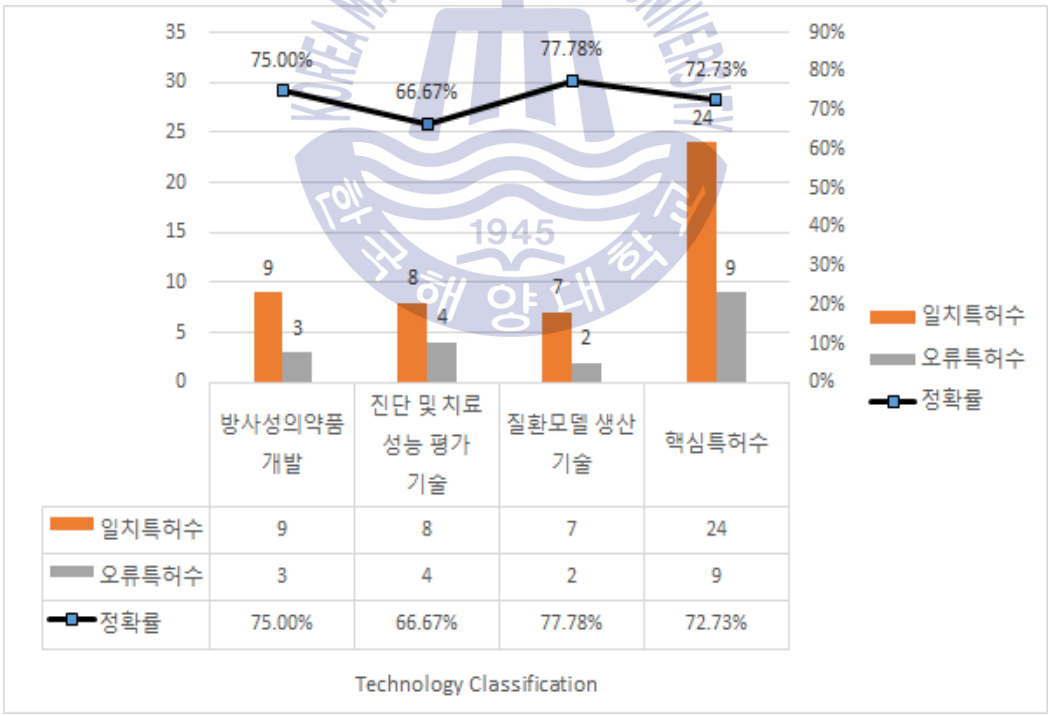


그림 4.22 두 번째 우선순위가중치 세트 적용 핵심 특허문헌의 정확률

Fig. 4.22 The accuracy of core patent documents applied 2nd priority weight set

진단 및 치료 성능 평가 기술 분야는 전체 12건 중에서 8건의 특허문헌이 일치하였고, 4건의 특허문헌이 불일치하여 정확률이 66.67%이다. 질환모델 생산기술 분야는 9건 중에서 7건의 특허문헌이 일치하였고, 2건의 특허문헌이 불일치하여 77.78%의 정확률을 나타내고 있다. 그리고 전체 33건의 핵심 특허문헌 중에서 24건의 특허문헌이 일치하고, 9건의 특허문헌이 불일치하여 72.73%의 정확률을 보이고 있다.

다음으로 특허조사원이 수작업을 통해 추출한 핵심 특허문헌의 결과를 대상으로 세 번째 우선순위가중치 세트를 적용하여 추출시스템을 통한 추출한 핵심 특허문헌의 중 일치하는 특허문헌과 오류특허문헌의 수 및 핵심 특허문헌 추출의 세부기술별 정확률은 그림 4.23과 같다.



	방사성의약품 개발	진단 및 치료 성능 평가 기술	질환모델 생산 기술	핵심특허수
일치특허수	9	8	7	24
오류특허수	3	4	2	9
정확률	75.00%	66.67%	77.78%	72.73%

그림 4.23 세 번째 우선순위가중치 세트 적용 핵심 특허문헌의 정확률

Fig. 4.23 The accuracy of core patent documents applied 3rd priority weight set

세 번째 우선순위가중치 세트를 적용한 결과, 방사선의약품 개발 분야는 전체 12건 중에서 9건의 특허문헌이 일치하였고, 3건의 특허문헌이 불일치하여 75%의 정확률을 보이고 있다. 진단 및 치료 성능 평가 기술 분야는 전체 12건 중에서 8건의 특허문헌이 일치하였고, 4건의 특허문헌이 불일치하여 정확률이 66.67%이다.

질환모델 생산기술 분야는 9건 중에서 7건의 특허문헌이 일치하였고, 2건의 특허문헌이 불일치하여 77.78%의 정확률을 나타내고 있다. 그리고 전체 33건의 핵심 특허문헌 중에서 24건의 특허문헌이 일치하고, 9건의 특허문헌이 불일치하여 72.73%의 정확률을 보이고 있다.

마지막으로 특허조사원이 수작업을 통해 추출한 핵심 특허문헌의 결과를 기준으로, 그림 4.24는 네 번째 우선순위가중치 세트를 적용된 핵심 특허문헌 중 일치하는 특허문헌과 오류특허문헌의 수 및 핵심 특허문헌 추출의 세부기술별 정확률을 보여준다. 네 번째 우선순위가중치 세트를 적용하여 세부기술별로 핵심 특허문헌을 추출한 결과, 방사선의약품 개발 분야는 전체 12건 중에서 9건의 특허문헌이 일치하였고, 3건의 특허문헌이 불일치하여 75%의 정확률을 보이고 있다.

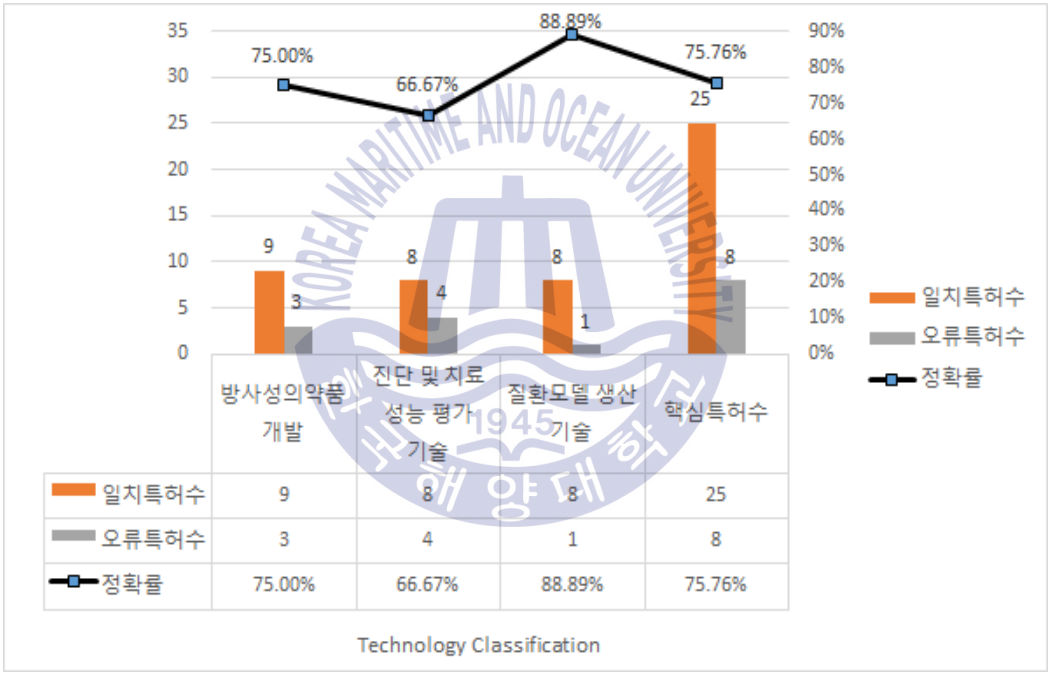


그림 4.24 네 번째 우선순위가중치 세트 적용 핵심 특허문헌의 정확률

Fig. 4.24 The accuracy of core patent documents applied 4th priority weight set

진단 및 치료 성능 평가 기술 분야는 전체 12건 중에서 8건의 특허문헌이 일치하였고, 4건의 특허문헌이 불일치하여 정확률이 66.67%이다. 질환모델 생산기술 분야는 9건 중에서 8건의 특허문헌이 일치하였고, 1건의 특허문헌이 불일치하여 88.89%의 정확률을 나타내고 있다. 그리고 전체 33건의 핵심 특허문헌 중에서 25건의 특허문헌이 일치하고, 8건의 특허문헌이 불일치하여 75.76%의 정확률을 보이고 있다.

특허조사원의 수작업을 통하여 추출한 핵심 특허문헌을 기준으로 추출시스템을 통해 추출된 핵심 특허문헌에 대하여 우선순위가중치 각 세트별 정확률은 그림 4.25와 같이 정리된다. 전체적으로 낮은 정확률을 나타내고 있으나, 그 중에서 네 번째 세트의 항목별가중치가 가장 높은 정확률인 75.76%를 보이고 있다.

이와 같이 추출시스템을 이용하여 핵심 특허문헌을 추출하는데 있어 정확률이 상대적으로 낮은 경향이 나타나고 있다. 하지만 이는 의뢰기관의 특허문헌을 핵심 특허로 반영한 결과로 앞으로 좀 더 명확한 선정 기준이 필요한 것으로 판단된다.

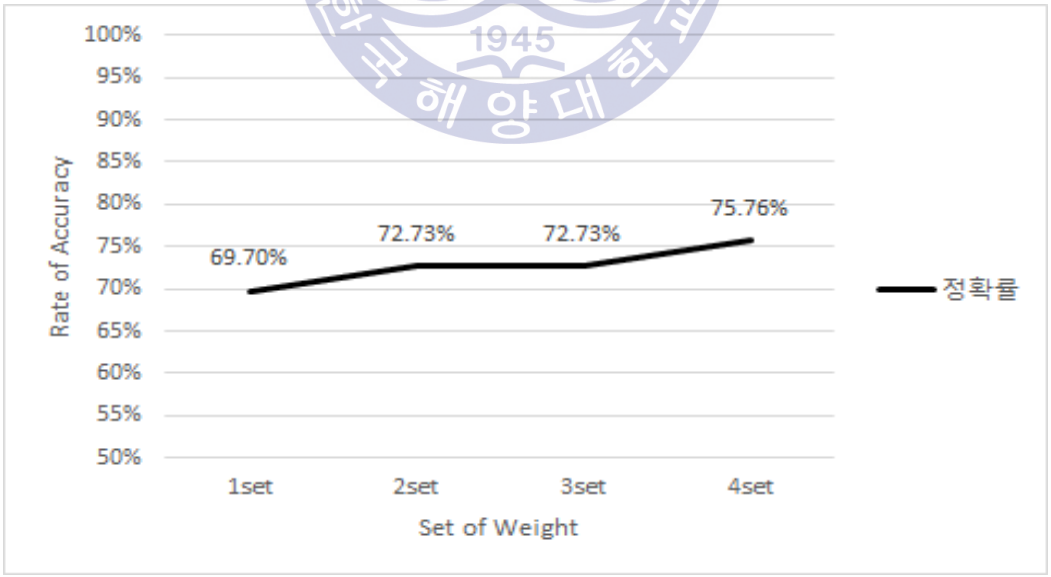


그림 4.25 우선순위가중치 세트별 핵심 특허문헌의 정확률

Fig. 4.25 The accuracies of core patent documents by sets of priority weight

4.6 결과 분석

추출시스템을 통하여 특허동향분석의 절차 중 기술요약서로부터 색인후보어를 추출하여 수작업을 통해 검색식을 작성하고, 특허검색서비스를 통해 추출한 후보 특허문헌 목록으로부터 유효 특허문헌 추출하고, 실험을 위해 수작업을 통해 추출된 유효 특허문헌을 이용하여 세부기술별 분류하며, 세부기술별로 분류된 유효 특허문헌으로부터 핵심 특허문헌을 추출하였다. 추출시스템의 성능을 확인하기 위하여 각 절차의 수행시간 및 정확률을 특허조사원이 수작업으로 수행하였던 결과와 비교하였다.

유효 특허문헌의 추출에 있어서 6개의 항목별가중치 세트를 적용한 추출시스템과 특허조사원이 수작업을 통해 추출한 1631건의 유효 특허문헌과 비교한 결과, 네 번째 세트의 항목별 가중치가 1563건의 유효 특허문헌으로 추출되어 가장 적은 차이를 보였으며, 정확율에 있어서도 86.88%로 비교적 높은 정확률을 보였다. 다만 각 국가별 정확율을 전체적으로 비교하면 한국을 비롯한 다른 나라들에 비해 일본의 경우, 정확률이 낮았는데 이는 특허검색서비스에서 제공하는 기계번역을 이용한 한국어 문장에 있어서 색인후보어와 번역된 한국어간의 차이로 인한 문제가 있음을 확인하였다.

그리고 추출시스템을 이용하여 유효 특허문헌 추출의 수행시간을 통한 성능을 비교하면, 추출시스템의 경우 후보 특허문헌 4036건에서 유효 특허문헌 1563건 추출을 수행하는 데는 약 36분 정도가 소요되어 특허조사원이 수작업을 통해 유효 특허문헌 1613건을 추출하는데 요구되었던 10,343분에 비해 약 287배 빠른 성능을 나타내었다.

추출시스템을 이용하여 방사성의약품 이용기술에 대한 1613건의 유효 특허문헌을 대상으로 세부기술별 분류하는데 있어서, 제안된 다섯 계층 신경망 구조를 통해 3개의 기술분류인 방사성의약품 개발, 진단 및 치료 성능 평가 기술, 질환모델 생산 기술로 분류하는데 성능을 비교하였다. 수행시간의 측면에서는 추출시스템이 약 73분이 필요하여 특허조사원이 수작업을 통해 수행하는 2,243분에 비해 약 31배 향상된 성능을 보였다.

그리고 특허조사원이 수작업을 통해 분류한 유효 특허문헌을 기준으로 추출시스템의 분류 결과의 정확률 측면에서는 방사성의약품 개발은 91.04%, 진단 및 치료 성능 평가 기술은 89.97%, 질환모델 생산 기술은 93.35%로 전체 91.08%의 정확률을 보였으며, 비교적 기술분류와 관련된 색인후보어의 영향이 명확한 질환모델 생산 기술 분야가 높은 정확률을 나타내고 있다. 또한 특허문헌을 대상으로 한 다른 알고리즘의 정확률과 비교하였을 때에도 신경망 알고리즘이 비교적 향상된 성능을 나타내고 있었다.

추출시스템을 이용한 핵심 특허문헌의 추출에 있어서도 특허조사원이 수작업을 통해 수행하는데 필요한 시간이 509분이었으나 추출시스템의 선정값 연산을 통해 수행한 시간은 약 25분으로 약 21배나 높은 성능을 보였다. 그리고 특허조사원이 추출한 핵심 특허를 기준으로 핵심 특허문헌의 추출에 있어 4개 우선순위가중치 세트들을 적용한 결과 네 번째 세트가 75.76%로 가장 높은 정확률 성능을 나타내었다. 다만 특허조사원이 추출한 핵심 특허와의 정확률에 있어서는 상대적으로 낮은 결과이지만, 이는 특허동향분석의 의뢰기관의 특허문헌을 핵심 특허문헌으로 반영한 결과로 앞으로 좀 더 명확한 선정 기준이 필요한 것으로 판단된다.

제 5 장 결론 및 향후 과제

본 논문에서는 특허동향분석의 절차 중 종래 특허조사원의 수작업에 의해 이루어지던 유효 특허문헌의 추출, 세부기술별 분류 및 핵심 특허문헌 추출 과정을 자동적으로 수행하는 핵심 특허문헌 추출시스템을 구현하였다. 또한 방사선의약품 이용기술에 관련된 특허문헌에 대하여 실험하고, 검증하였다.

핵심 특허문헌 추출시스템을 구현하기 위하여 유효 특허문헌의 추출은 기술요약서로부터 검출한 검색식으로 검색한 후보 특허문헌 목록으로부터 불리언검색의 문제점인 연구개발기술과 관련 없거나 적은 특허문헌을 제거하기 위하여 코사인 유사도를 이용하여 추출하였다. 유사도는 색인후보어의 TF-IDF가중치를 이용한 기술요약서의 문서벡터와 항목별가중치, TF-IDF가중치 및 비색인어가중치를 반영한 각 특허문헌의 문서벡터간의 코사인 값을 이용하여 측정하였다.

세부기술별 분류는 신경망 알고리즘을 이용하여 한 개의 입력층과 세 개의 은닉층, 한 개의 출력층으로 다섯 계층의 신경망의 구조로 이루어지며, 입력변수값은 색인후보어의 TF-IDF가중치이고, 제1은닉층에 입력값에 기술분류와 관련된 색인후보어에 대하여 분류가중치를 제공하여 분류하였다. 핵심 특허문헌의 추출은 분류된 세부기술별로 각 유효 특허문헌의 항목에 대한 우선순위를 정하고, 항목에 기재된 값과 우선순위에 따른 우선순위가중치를 연산한 선정값을 순위화하여 특허조사원이 지정하였던 수의 핵심 특허문헌을 추출하였다.

핵심 특허문헌 추출시스템의 실험은 방사성의약품 이용기술에 관한 기술요약서로부터 작성한 검색식을 이용하여 특허검색서비스로부터 검색한 4306건의 후보 특허문헌을 대상으로 실험하였다. 핵심 특허문헌 추출시스템의 검증은 기존의 특허조사원이 수작업을 통해 수행한 유효 특허문헌, 세부기술별 분류 및 핵심 특허문헌의 결과와 정확성 및 수행시간을 비교하였다.

유효 특허문헌의 추출과 관련하여 정확률에 있어서는 86.9%로 일본 특허문헌을 제외하면 수작업과 비슷한 좋은 결과를 얻었다. 그리고 수행시간에서는 수작업의 경우 10,343분의 시간이 필요하였으나, 핵심 특허문헌 추출시스템은 약 36분으로 높은 성능을 보이고 있다.

세부기술별 분류에 있어서는 특허조사원이 수행하였던 분류와 동일하게 방사성의약품 이용기술을 방사성의약품 개발기술, 진단 및 치료 성능 평가 기술, 질환모델 생산 기술로 분류하였다. 핵심 특허문헌 추출시스템의 정확률은 91.08%로 다른 알고리즘에 비해 향상된 성과를 얻었다. 그리고 세부기술별 분류의 수행시간에 있어서는 특허조사원이 수작업으로 분류하는데 2,243분이 걸린데 비해, 핵심 특허문헌 추출시스템은 약 73분으로 31배가 넘는 효과가 있었다.

핵심 특허문헌의 추출은 방사성의약품 개발, 진단 및 치료 성능 평가 기술 및 질환모델 생산 기술로 분류된 유효 특허문헌에 대하여 선정값을 기준으로 핵심특허를 추출한 결과는 특허조사원이 수행한 결과를 기준으로 75.76%의 정확률을 보였다. 하지만 이는 특허조사원이 수작업으로 핵심 특허문헌을 추출하는데 있어 의뢰기관의 특허가 반영된 결과로 보다 명확한 선정기준의 보완이 요구된다. 그리고 핵심 특허문헌의 추출의 수행시간은 특허조사원이 핵심 특허문헌을 추출하는데 509분의 시간이 요구된 반면, 추출시스템은 약 25분으로 빠른 성능을 나타냈다.

전체적으로 특허조사원의 수작업은 총 13,095분의 시간이 필요하였으나, 제안된 추출시스템은 134분으로 약 97배 이상의 향상된 성능결과를 도출하였다. 따라서 제안된 시스템은 특허조사원에게 특허동향분석에 있어 발생하는 시간적인 한계를 해결하고, 오류의 발생 가능성을 감소시키는 효과를 제공하고 있다.

향후에는 최근 활발히 연구되고 있는 딥러닝 알고리즘을 이용하여 보다 풍부한 학습 데이터를 근간으로 학습을 수행하여 정확률에 있어 성능을 향상시키고자 한다. 그리고 방사성의약품 이용기술 외에 다른 연구개발 기술의 특허동향분석의 유효 특허문헌 추출, 세부기술별 분류 및 핵심특허문헌 추출에도 제안된 시스템을 적용하고자 한다.

참고문헌

[1] 과학기술정책연구원, 신성장동력 사업과 지역 혁신 사업의 연계 전략, 2011.

[2] 미래창조과학부 한국과학기술기획평가원, 2015년도 국가연구개발사업 투자현황 보고서, 2016.

[3] 고용수, 김치용, 김성수, 정부의 연구개발(R&D) 투자 규모 및 향후 투자방향에 대한 제언, 한국과학기술기획평가원, 2012.

[4] 특허청 제16회 국가과학기술 위원회, 국가연구개발사업 효율화를 위한 특허정보 활용확산 계획(안)」, 2004.

[5] 국가법령정보센터, 미래창조과학부, 국가연구개발사업의 관리 등에 관한 규정, 2005.

[6] 강귀용, 특허동향분석(PTA) 기법을 통한 미래지속가능 기술 예측에 관한 연구, 경기대학교 대학원, 박사학위 논문, 2015.

[7] 김갑조, 주제기반 특허분석을 통한 기술예측시스템 개발에 관한 연구, 고려대학교 대학원, 박사학위 논문, 2014.

[8] 한국저작권위원회, 선행기술의 조사와 지재권 관리 전략, 2007.

[9] 강승식, 한국어 형태소 분석과 정보검색, 홍릉과학출판사, 2002.

[10] 이운재, 김선배, 김길연, 최기선, “모듈화된 형태소 분석기의 구현”, 한글 및 한국어 정보처리 학술대회 발표논문집, pp. 123-136, 1999.

[11] 강승식, 이하규, “한국어 형태소 분석기 HAM의 형태소 분석 및 철자 검사 기능”, 한글 및 한국어 정보처리 학술대회 발표논문집, pp. 216-252, 1996.

[12] Shin, J.S. and Lee, C.H., “Artificial intelligence : Automatic classification of documents using word correlation”, Journal of Korea Information Processing Society, vol. 6, no. 9, pp. 2422-2430, 1999.

[13] 이정훈, 전서현, “연관 관계와 TF*IDF를 이용한 검색 결과 Re-Ranking”, 한국정보과학회 한국 컴퓨터 종합학술대회 발표논문집, 제37권, 제1C호, pp. 349-352, 2010.

[14] Go, G.S., Jung, W.G., Shin, Y.G., Park S.S. and Jang, D.S., “A study on the patent information retrieval algorithm”, Proceedings of the Korean Society of Computer Information, vol. 19, no.2, pp. 369-371, 2011.

[15] Burke, R., “Recommender systems : An introduction”, International Journal of Human-Computer Interaction, vol. 28, no. 1, pp. 72-73, 2012.

[16] 원상훈, 노태길, 손기준, 박정희, 이상조, “특허정보 검색을 위한 벡터스페이스 검색 모델의 적용”, 한국정보과학회 2003년도 봄 학술대회 발표논문집, 제30권, 제1호, pp. 516-518, 2003.

[17] Blanchard, A., “Understanding and customizing stopword lists for enhanced patent mapping”, Journal of World Patent Information, vol. 29, no. 4, pp. 308-316, 2007.

[18] 정영미, 임혜영, “SVM 분류기를 이용한 문서 범주화 연구”, 한국정보관리학회 논문지, 제17권, 제4호, pp. 229-248, 2000.

[19] 강윤희, 박용범, “SVM을 이용한 기술정보 문서의 디렉터리 기반 문서 분류 시스템의 설계 및 구현”, 한국인터넷방송통신학회 논문지, 제4권, 제1호, pp. 71-78, 2004.

[20] 김혜숙, 박상철, 김수형, "단어가중치 기반 문서간 유사도 측정에 관한 연구", 한국멀티미디어학회 학술대회 발표논문집, 제2003권, 제1호, pp. 198-201, 2003.

[21] 박태정, SVM 유사도분석을 이용한 웹 컨텐츠 복제 추정에 관한 연구, 서강대학교 대학원, 석사학위 논문, 2005.

[22] Ahn, J., "Similarity measuring method for essential patent technology using text mining", Proceedings of the Korean Computer Conference, vol. 36, no. 1, pp. 1-5, 2009.

[23] Yang, B.J. and Shim, J.H., "Practical datasets for similarity measures and their threshold values", Journal of Society for e-Business Studies, vol. 18, no. 1, pp. 97-105, 2013.

[24] Kim, M.J. and Lee, S.J., "Measures of abnormal user activities in online comments based on cosine similarity", Journal of Korea Institute of Information Security and Cryptology, vol. 24, no. 2, pp. 335-343, 2014.

[25] Lee, D.J. and Shim, J.H., "Survey on vector similarity measures : Focusing on algebraic characteristic", Journal of Society for e-Business Studies, vol. 17, no. 4, pp. 209-219, 2012.

[26] Lee, D.J., Park, J.H., Shim, J.H. and Lee, S.G., "An efficient similarity join algorithm with cosine similarity predicate", Proceedings of the Lecture Notes in Computer Science, vol. 2, no. 6262, pp. 422-436, 2010.

[27] Helen J., *et al.*, "The limitation of term coocurrence data for query expansion in document retrieval system", Journal of American Society for information Science, vol. 24, no. 5, pp. 378-383, 1991.

[28] Qiu, Y. and Frei, H.P., "Concept based query expansion", Proceedings of the 16th ACM SIGIR Conference on Research and development in Information Retrieval, vol. 27, pp. 160-169, 1993.

[29] Pabitra, M., Murthy, C.A. and Pal, S.K., "Unsupervised feature selection using feature similarity", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 3, pp. 301-312, 2002.

[30] Song, Y. and Hou, L., "A methodology for text classification based on feature clustering", Proceedings of the International Conference on Automatic Control and Artificial Intelligence, pp. 119-124, 2012.

[31] Lin, Y.S., Jiang, J.Y. and Lee S.J., "A similarity measure for text classification and clustering", IEEE Transactions on Knowledge and Data Engineering, vol. 26 no. 7, pp. 1575-1590, 2014.

[32] Tang, J., *et al.*, "Patent miner : Topic-driven patent analysis and mining", Proceedings of the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 1366-1374, 2012.

[33] Zhang, *et al.* "Patentline : Analyzing technology evolution on multi-view patent graphs", Proceedings of the 37th ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1095-1098, 2014.

[34] Lee, R.C., Lee, W.K., Song, J.S. and Eom, C.E., "Measuring patent similarity based meta-path in heterogeneous network", Proceedings of the Conference on Korea Information Science Society, vol. 2015, no. 6, pp. 724-726, 2015.

[35] Rabin, M.O., Fingerprinting by Random Polynomials, Department of Mathematics The Hebrew University of Jerusalem, 1981.

[36] Mühleisen, H., Walther, T. and Tolksdorf, R., "Multi-level indexing in a distributed self-organized storage system", IEEE Transactions on Evolutionary Computation, pp. 989-994, 2011.

[37] Chowdhury, G.G. and Sudatta, C., Introduction to Digital Libraries, Facet Publishing, 2002.

[38] Kim, J.H., Kim, Y.J. and Kim, J.B., "A study on similarity analysis of national R&D programs using R&D project's technical classification", Journal of Digital Contents Society, vol. 13, no. 3, pp. 317-324, 2012.

[39] Kang, J.S., Lee, H.J. and Moon, Y.H., Apparatus and method for configuring a comprehensive intellectual property rights star network by detecting patent similarity, Korea Patent 1009058920000, July 26, 2009.

[40] 김한경, 나휘동, 이금희, 이종혁, "문장구조 유사도와 단어 유사도를 이용한 클러스터링 기반의 통계기계번역", 정보과학회 논문지 : 소프트웨어 및 응용, 제37권, 제4호, pp. 297-304, 2010.

[41] Koutroumbas, K. and Kalouptsidis, N., "Nearest neighbor pattern classification neural networks", Proceedings of the IEEE International Conference on Neural Networks, vol. 5, pp. 2911-2915, 1995.

[42] Yang, Y. and Chute, C.G., "An example-based mapping method for text categorization and retrieval", ACM Transactions on Information Systems, vol. 12, no. 3, p. 252, 1994.

[43] Mitchell, T.M., Machine Learning, McGraw-Hill, 1997.

[44] Vapnik, V. and Cortes, C., "Support-vector networks", Journal of Machine Learning, vol. 20, no. 3, pp. 273-297, 1995.

[45] Wajeed, M.A. and Adilakshmi, T., "Text classification using machine learning", Journal of Theoretical and Applied Information Technology, vol. 7, no. 2, pp. 119-123, 2009.

[46] Jing, H., *et al.*, "Semantic naïve bayes classifier for document classification", International Joint Conference on Natural Language Processing, pp. 1117-1123, 2013.

[47] United States Patent and Trademark Office, https://www.uspto.gov/web/patents/classification/selectnumwithtitle.htm, Accessed November 15, 2016.

[48] Japan Patent Office, https://www.jpo.go.jp/torikumi_e/searchportal_e/pdf/classification/fi_f-term.pdf, Accessed November 15, 2016.

[49] European CLAssification System, https://worldwide.espacenet.com/help?topic=ecla&method=handleHelpTopic&locale=en_ep, Accessed November 15, 2016.

[50] Cooperative Patent Classification, https://www.uspto.gov/patents-application-process/patent-search/classification-standards-and-development, Accessed November 15, 2016.

[51] WIPO, http://www.wipo.int/classifications/ipc/en/, Accessed November 15, 2016.

[52] 한국특허정보원, http://www.kipi.or.kr/kipi/, Accessed November 16, 2016.

[53] 박찬정, 신수봉, 정병일, 진성희, "캠퍼스 특허전략 유니버시아드 교육 프로그램 개발에 관한 연구 : 인하대학교 사례를 중심으로", 한국지식재산교육연구학회 2013 추계학술대회 발표논문집, pp. 153-158, 2013.

[54] 박찬정, “비특허 문헌을 활용한 캠퍼스 특허전략 유니버시아드 기술 접근 방법”, 한국지식재산교육연구학회 2014 춘계학술대회 발표논문집, pp. 285-288, 2014.

[55] Larkey, L.S., “A patent search and classification system”, Proceedings of the 4th ACM SIGIR Conference on Digital Libraries, pp. 179-187, 1999.

[56] Littlestone, N., Machine Learning 2, Kluwer Academic Publishers, 1988.

[57] Ioana, C., Bot, R.I. and Wanka, G., Patent document classification based on mutual information feature selection, Technische Universität Chemnitz. Fakultät für Mathematik, 2004.

[58] Marc, K. and Zacca, F., “Automatic categorization applications at the european patent office”, Proceedings of the World Patent Information, vol. 24, no. 3, pp. 187-196, 2002.

[59] Fall, C.J., *et al.*, “Automated categorization in the international patent classification”, ACM SIGIR Forum, vol. 37, no. 1, pp. 10-25, 2003.

[60] Fall, C.J., *et al.*, “Computer-assisted categorization of patent documents in the international patent classification”, Proceedings of the International Conference on Chemical Information, 2003.

[61] Tong, L.H., He, C. and Shen, L., "Automatic classification of patent documents for TRIZ users", Proceedings of the World Patent Information, vol. 28, no. 1, pp. 6-13, 2006.

[62] Trappey, A.J., *et al.*, “Development of a patent document classification and search platform using a backpropagation network", Journal of Expert Systems with Applications, vol. 31, no. 4, pp. 755-765, 2006.

[63] Mathiassen, H.R. and Daniel, O.A., “Automatic categorization of patent applications using classifier combinations”, Proceedings of the Lecture Notes in Computer Science, no. 4224, pp. 1039-1047, 2006.

[64] Wang, W., Li, S. and Wang. C., “ICL at NTCIR-7 : An improved KNN algorithm for text categorization”, Proceedings of the NTCIR-7 Workshop Meeting, 2008.

[65] Xiao, T., *et al.*, “KNN and re-ranking models for english patent mining at NTCIR-7”, Proceedings of the NTCIR-7 Workshop Meeting, 2008.

[66] Wu, C.H., Yun, K. and Tao, H., “Patent classification system using a new hybrid genetic algorithm support vector machine”, Journal of Applied Soft Computing, vol. 10, no. 4, pp. 1164-1177, 2010.

[67] Lee, J.A., Seo, H.K. and Han, K.Y., “Refined IPC classification system based on KNN using patent search results”, Proceedings of Conference on the Korea Information Science Society, vol. 38, no. 2A, pp. 256-259, 2011.

[68] Khattak, A.S. and Heyer, G., "Significance of low frequent terms in patent classification using IPC hierarchy", Proceedings of the 6th International Multi-Conference on Computing in the Global Information Technology, pp. 239-250, 2011.

[69] 이재안, 서형국, 한규열, “특허 문서 검색 결과를 이용한 KNN 기반의 특허 분류 시스템”, 한국정보과학회 2011 가을학술대회 발표논문집, 제38권, 제2호, pp. 256-259, 2011.

[70] Chen, Y.L. and Chang, Y.C., “A three phase method for patent classification”, Journal of Information Processing & Management, vol. 48, no. 6, pp. 1017-1030, 2012.

[71] Park, C.J., Seong, D.S. and Lee, K.B., “Automatic IPC classification for patent documents using machine learning”, Journal of Korean Institute of Information Technology, vol. 10, no. 4, pp. 119-128, 2012.

[72] Park, C.J., *et al.*, “Automatic IPC classification of patent documents using the term clustering”, Journal of Korean Institute of Information Technology, vol. 12, no. 9, pp. 127-139, 2014.

[73] 박찬정, 김기용, 성동수, “KNN을 이용한 융합기술 특허문서 자동 IPC 분류”, 한국정보기술학회 논문지, 제12권, 제3호, pp 175-185, 2014.

[74] Kang, J.H., *et al.*, “A comparative study on patent document classification algorithms”, Proceedings of the Conference on Korean Institute of Intelligent Systems, vol. 26, no. 1, pp. 9-10, 2016.

[75] Khattak, A.S. and Heyer, G., “Significance of low frequent words in patent classification”, Proceedings of the 11th International Conference on Innovative Internet Community Systems, pp. 8-13, 2011.

[76] 김태중, 이명선, 최호남, “특허 발명의 명칭에 쓰인 단어를 이용한 기술동향 분석 연구”, 한국콘텐츠학회 논문지, 제10권, 제4호, pp. 433-437, 2010.

[77] Kim, S.H. and Eom, J.E., “A study on the documents automatic classification using machine learning”, Journal of Information Management, vol. 39, no. 4, pp. 47-66, 2008.

[78] 김갑조, 박상성, 신영근, 정원교, 고광수, 장동식, “특허 분석을 위한 알고리즘에 관한 연구”, 대한산업공학회 춘계공동학술대회 발표논문집, pp. 1266-1270, 2011.

[79] 최성배, 김태완, “귀납학습 알고리즘을 이용한 문서 자동 분류 시스템 설계”, 인제논총, 제18권, 제1호, pp. 443-454, 2003.

[80] 엄재은, 기계학습을 이용한 문서 자동분류에 관한 연구, 중앙대학교 대학원, 석사학위 논문, 2008.

[81] Hinton, G.E., *et al.*, "Deep neural networks for acoustic modeling in speech recognition : The shared views of four research groups", IEEE Signal Processing Magazine, pp. 82-97, 2012.

[82] Yuan, Y.X., "Step-sizes for the gradient method", Journal of AMS/IP Studies in Advanced, vol. 42, no. 2, pp. 785-796, 2008.

[83] 유선희, 이용호, 원동규, "특허정보분석을 이용한 기술파급효과 측정에 관한 연구", 한국기술혁신학회 기술혁신학회지, 제10권, 제2호, pp. 687-705, 2007.

[84] 노경란, 한상완, "특허분석을 통한 과학기술자의 과학논문 인용행태에 관한 구현", 한국정보관리학회 정보관리학회지, 제23권, 제3호, pp. 223-239, 2006.

[85] Kando, N., "What shall we evaluate? preliminary discussion for the NTCIR patent IR challenge based on the brain storming with the specialized intermediaries in patent searching and patent attorneys", Proceedings of the ACM SIGIR 2000 Workshop on Patent Retrieval, 2000.

[86] 한국철도기술연구원, 광역급행버스의 좌석예약방법 및 상기 방법에 의한 광역급행버스의 좌석예약장치, 한국 특허번호 1014905180000, 2015년 1월 30일.

[87] David, D.L., *et al.*, "Rcv1 : A new bench mark collection for text categorization research", Journal of Machine Learning Research, vol. 5, pp. 361-397, 2004.

[88] Park, Y.J., "Social tagging-based recommendation platform for patented technology transfer", Journal of Intelligent Information Systems on Korea Intelligent Information Systems Society, vol. 21, no. 3, pp. 53-77, 2015.

[89] 국가과학기술위원회, 2011년도 국가연구개발사업 특허기술동향조사 추진계획(안), 2011.

[90] 한국지식재산전략원, http://www.kista.re.kr/, Accessed November 18, 2016.

[91] 안성만, “딥러닝의 모형과 응용사례”, 한국지능정보시스템학회 지능정보연구, 제22권, 제2호, pp. 127-142, 2016.

[92] Wisdomain, Inc., Wisdomain guide, 2016.

[93] 원상훈, 노태길, 손기준, “특허정보 검색을 위한 벡터스페이스 검색모델의 적용”, 한국정보과학회 학술대회 발표논문집, 제30권, 제1호, pp. 516-518, 2003.

[94] 법제처, 특허법 제64조 제1항, 2014.