韓國IT서비스學會誌
第17卷 第2號
2018年 6月, pp.129-142

Journal of Information Technology Services
https://doi.org/10.9716/KITS.2018.17.2.129

Word2Vec 학습을 통한 의미 기반 해외 유사 특허 검색 방안*

백민지** · 김남규***

Identifying Similar Overseas Patent Using Word2Vec-Based Semantic Text Analytics*

Minji Paek** · Namgyu Kim***

■ Abstract ■

Recently, the number of patent applications have been increasing rapidly every year as the importance of protecting intellectual property rights becomes more important. Patents must be inventive and have novelty. Especially, the novelty implies that the corresponding invention is not the same as the previous invention. To confirm the novelty, prior art search must be conducted before and after the application. The target of prior art search should include not only Korean patents but also foreign patents. Search of foreign patents should be supported by multilingual search techniques. However, a dictionary-based naive approach shows a limitation because some technical concepts are represented in different terms according to each nation. For example, a Korean term and a Japanese term may not be synonym even though they represent the same technical concept. In this paper, we propose a new method to map semantic similarity between technical terms in Korean patents and Japanese patents. To investigate different representations in each nation for the same technical concept, we identified and analyzed pairs of patents those are mutually connected with priority claim relationship. By performing an experiment with real-world data, we showed that our approach can reveal semantically similar technical terms in other language successfully.

Keyword : Patent Analysis, Word2Vec, Semantic Similarity, Multilingual Search, Prior Art Search

Submitted : January 31, 2018 1st Revision : March 3, 2018 Accepted : March 14, 2018

* 이 논문은 2017년 대한민국교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2017S1A5A2A03067632).

** 국민대학교 비즈니스IT전문대학원 석사과정
*** 국민대학교 경영정보학부 부교수, 교신저자

1. 서 론

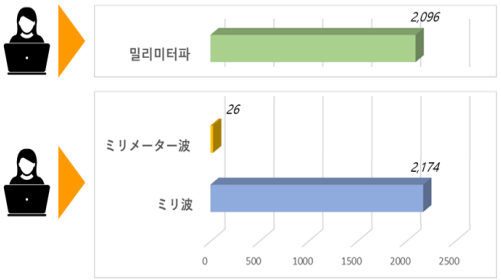
최근 여러 분야에서 새로운 기술이 고안되고 사용됨에 따라, 이들 기술을 독점적으로 관리하기 위한 지식 재산권에 대한 관심이 높아지고 있다. 지식 재산권은 산업 재산권과 저작권을 모두 포함하는 개념이며, 산업 재산권은 다시 특허권, 실용신안권, 상표권, 그리고 디자인권으로 분류(Chang, 2017)된다. 이 가운데 기술의 보호에 가장 널리 활용되는 특허의 경우, 보호받고자 하는 기술을 출원하여 특허권을 취득하게 되면 해당 권리에 대해 일정 기간 독점권을 갖는다. 즉 해당 기간 동안 제 3자는 특허권자의 동의 없이 해당 권리를 행사할 수 없으므로, 기업 간 그리고 국가 간 유망 기술에 대한 특허를 확보하기 위한 경쟁이 더욱 심화하고 있다.

세계지식재산권기구(WIPO)에서 발표한 2016년의 통계를 보면 2016년 PCT(국제특허협력조약)에 따른 특허 출원은 총 23만 3천 건에 달하며, 이는 전년도인 2015년 대비 7.3% 증가한 수치이다. 우리나라의 경우 전 세계 PCT 특허 출원의 약 6.7%에 해당하는 15,560건의 출원이 이루어졌으며, 출원 건수 기준 세계 5위를 차지할 정도로 특허 출원이 매우 활발하게 이루어지고 있다.

특허가 등록되기 위해서는 출원한 발명이 반드시 진보성과 신규성을 동시에 가져야 한다. 진보성이란 발명의 창작 수준의 난이도를 뜻하며, 해당 발명이 속하는 기술 분야에서 통상의 지식을 가진 사람이 용이하게 발명할 수 없는 정도의 창작 난이도를 갖추어야 발명의 진보성이 인정된다. 한편 신규성은 발명의 새로움을 뜻하며, 출원 시점을 기준으로 이미 공지된 다른 발명과 동일하지 않아야 발명의 신규성이 인정된다.

진보성과 신규성을 상실한 특허는 등록될 수 없기 때문에, 출원 전에 해당 발명과 유사한 기술이 존재하는지 여부를 확인하기 위한 선행기술조사(Prior Art Search)가 반드시 수행되어야 한다. 선행기술조사는 특허 기술을 분류하기 위해 부여된 다양한 코드를 기반으로 검색을 수행하는 코드 검색 방식과, 핵심 기술을 키워드로 사용하여 검색하는 키워드 검색 방식의 두 가지로 진행된다. 특허의 경우 타 분야에 비해 코드 체계가 매우 잘 정리되어 있으므로, 일반적으로 코드 기반 검색을 통해 특정 발명이 속한 기술 분야의 범주를 효과적으로 한정할 수 있다. 하지만 동일 코드 내에 속한 발명이 매우 많기 때문에, 해당되는 범주 내에서 유사 발명을 찾기 위해서는 다시 키워드 기반 검색을 반복적으로 수행하게 된다.

선행기술조사의 대상은 국내에서 출원된 특허뿐만 아니라 해외에서 출원된 특허도 포함한다. 코드 기반 검색의 경우 국제특허분류(IPC)를 통해 국내 특허의 검색과 유사한 방식으로 수행할 수 있다. 하지만 키워드 기반 검색의 경우 해당 국가의 언어로 검색이 이루어져야 하므로, 적절한 다국어 검색 기술의 활용이 매우 중요하다. 다국어 키워드 검색은 일반적으로 동의어 사전을 통해 수행된다. 예를 들어 한국어 입력을 통해 일본어로 출원된 발명을 검색하고자 하는 경우, “밀리미터파”를 입력하면 이 값이 동의어 사전을 통해 “ミリメーター波”로 번역되어 일본 특허 검색 시스템으로 전달되는 방식이다. 하지만 이러한 사전 기반 다국어 검색은 동일한 개념이 국가 간 상이한 언어로 표현되는 경우 검색에 한계를 나타낸다(<Figure 1> 참조).



〈Figure 1〉 Differences Terms for the Same Concept

<Figure 1>은 1~10mm의 파장과 30~300GHz의 주파수를 갖는 전파인 “밀리미터파”에 대한 한국과 일본 발명의 검색 예이다. 한국 특허 검색 시스템에서 “밀리미터파”의 검색어로 2,096건의 특

허가 조회된 반면, 일본 특허 검색 시스템에서 “밀리미터파”의 일본어 번역인 “ミリメーター波”의 검색어로는 불과 26건이 조회되었다. 이는 일본에서는 한국의 “밀리미터파”에 해당하는 용어로 단순 동의어인 “ミリメーター波”가 아닌 “ミリ波”를 주로 사용하기 때문이며, 실제로 검색어를 “ミリ波”로 수정 검색하여 2,174건의 조회 결과를 얻을 수 있었다. 이 예는 단순히 동의어 사전에 의지하여 국가별 검색을 수행하는 경우, 동일한 기술 개념이 국가별로 상이한 어휘로 표현될 때 관련 선행기술을 충분하게 찾아내지 못한다는 한계를 보인다. 이러한 현상은 새로운 기술 개념이 빠르게 출현하며, 이들 개념에 대한 용어의 표준화가 충분히 이루어지기도 전에 다수의 특허가 출원되는 최근 환경에서 더욱 큰 한계로 나타날 수 있다.

이러한 한계를 극복하기 위해 본 연구에서는 실제 출원이 이루어진 두 국가의 특허 문서로부터 기술용어의 용례를 파악하고, 이를 통해 양 국가의 기술용어 간 의미적 유사성에 기반을 둔 매핑을 수행한다. 제안 방법론은 실제 용례로부터 용어 간 의미적 유사성을 도출하기 위해, 동일한 내용을 상이한 언어로 기술한 문서 쌍에 대해 Word2Vec 알고리즘을 적용한다. 이러한 조건을 만족시키는 문서 집합으로, 본 연구에서는 특허 제도 중 하나인 우선권(Priority Claim) 제도로 연결된 한국과 일본의 특허 문서를 활용한다. 우선권으로 연결된 한국과 일본의 두 특허는, 서로 표기된 언어는 다르지만 특허 명세서의 내용은 거의 동일한 것으로 간주된다. 즉 본 연구에서는 우선권으로 연결된 한국과 일본의 특허 문서에 대한 학습을 통해 양국 기술용어 간 유사성을 도출하고, 이를 기반으로 다국어 특허 검색을 수행할 수 있는 방안을 제시한다. 다국어 특허 검색을 통해 정적인 사전을 통한 단어 검색의 한계점을 보완할 수 있으며, 최신 기술 동향에 맞는 실사용 용어를 검색할 수 있고 그로 인해 선행기술조사 시, 양질의 선행기술을 풍부하게 찾아낼 수 있다.

본 논문의 이후 구성은 다음과 같다. 다음 장인 제 2장에서는 본 연구와 관련된 기존 연구들을 요약하고, 제 3장에서는 본 연구에서 제안하는 방법론을 소개한다. 다음 장인 제 4장에서는 제안 방법론을 실제 데이터에 적용한 실험 결과를 살펴보고, 마지막 제 5장에서는 본 연구의 기여 및 한계를 요약한다.

2. 선행 연구

기술의 권리를 획득하기 위해 많은 기업들이 특허를 출원하고 있으며, 이와 비례하여 특허 관련 연구들도 매우 활발하게 진행되고 있다.

또한 최근에는 방대한 텍스트 데이터가 실시간으로 생성됨에 따라 텍스트 분석을 다루는 연구(Lee et al., 2017; Wong and Kim, 2016)가 다방면에서 이루어지고 있다. 특히 텍스트 분석에 활용되는 데이터 중 특허 문서는 정형 데이터 형태의 다양한 서지정보 뿐 아니라 비정형 텍스트 데이터를 포함하고 있으며, 특허의 텍스트는 다른 문서에 비해 매우 정제되어 있기 때문에 여러 목적의 텍스트 분석 실험에 사용되기 적합한 특징을 가지고 있다. 이러한 특징으로 인해 텍스트 분석을 다루는 많은 연구의 실험 데이터로 활용되고 있다. 이러한 연구의 대표적인 예로는 특허 분석을 통해 기술경쟁력을 분석한 연구(Park and Kim, 2007), IT 유망기술 도출에 관한 연구(Kim and Hwang, 2009), 기술의 개발 동향 및 연구 동향 파악을 위한 연구(Kim, 2009; Lee et al., 2008; Moon et al., 2006; Park and Kim, 2009; Park and Kim, 2010), 특허가 관심 연구 분야에 미치는 영향에 대한 연구(Yoon et al., 2010), 특허 분석 결과를 통한 기술 예측 연구(Jun, 2011), 과학기술의 지식흐름을 분석한 연구(Park and Sung, 2008) 등이 있다.

특허 데이터를 실험 데이터로 활용하여 타분야의 이슈를 해결하기 위한 연구가 매우 활발하게 수행된 것에 비해, 특허 데이터 분석을 통해 특허 관련 프로세스 자체를 개선하기 위한 연구는 상대적으로 매우 부족한 현실이다. 대표적인 연구로는

분쟁특허 간 유사도를 측정한 연구(Bergmann et al., 2008), 효율적 특허 검색에 관한 연구(Zhang et al., 2013) 등이 있으며, 단어 간 의미적 유사성을 계산하여 대체어의 후보를 추출하는 방법을 제안한 연구(Baik et al., 2009) 등이 있다. 전술한 바와 같이 특허 데이터 분석을 통해 특허 프로세스를 개선하기 위한 시도는 활발하게 이루어지지 않았으며, 다른 언어로 작성된 다국어 특허 문서의 검색을 위한 연구는 특히 찾아보기 어렵다.

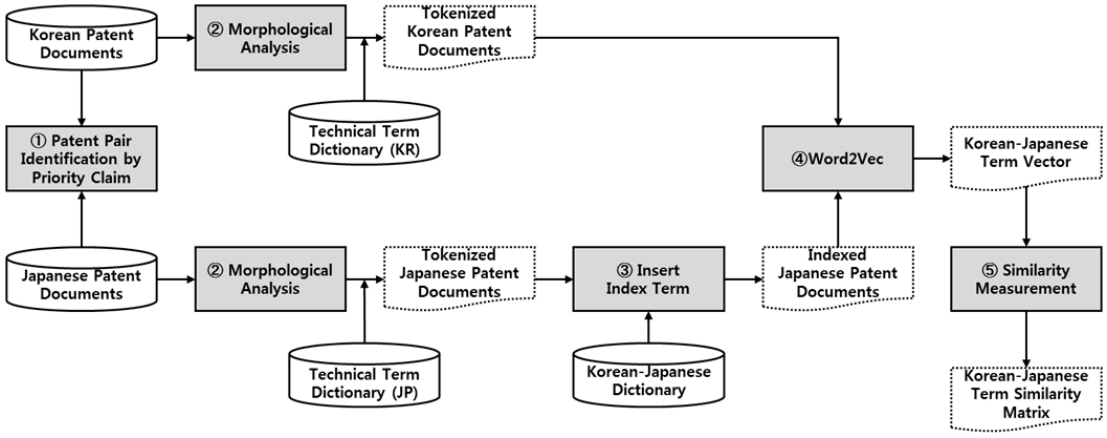
특허는 지금까지 없었던 새로운 기술에 대하여 진보성과 신규성이 판단되는 경우에 취득할 수 있으므로, 유사 특허를 찾기 위한 선행기술조사 과정은 특허 출원의 핵심 과정이다. 또한 선행기술조사는 국내 특허뿐 아니라 해외에서 출원된 특허도 포함하여 수행되므로, 이 과정에 매우 많은 시간이 소요될 뿐 아니라 검색 정확도 측면에서의 한계도 나타난다. 즉 다국어 검색의 경우 일반적으로 단순 직역을 통해 검색이 이루어지지만, 동일한 기술 개념에 대해 국가별로 상이한 용어를 사용하는 경우 검색의 정확도는 현저히 낮아지게 된다. 따라서 본 연구에서는 다국어 특허 문서에 대해 Word2Vec 학습 기반 의미적 유사성 분석(Mikolv et al., 2013)을 수행하고, 이를 통해 해외 유사 특허 검색의 정확도를 향상시키기 위한 방안을 제시한다.

3. 제안 방법론

3.1 연구 모형

본 절에서는 제안 방법론의 전체 개요를 간략히 소개한다. <Figure 2>는 제안 방법론의 과정을 도식화한 것으로, 그림에서 데이터는 원통형, 프로세스는 직사각형, 그리고 프로세스에 의한 산출물은 점선 도형으로 표시되어 있다.

우선 내용의 유사성이 매우 높은 한국과 일본의 특허 문서 쌍을 확보하기 위해, 분석 대상 한국 특허에 대해 우선권으로 연결된 일본 특허 문서를 수집한다(①). 다음으로 양국의 특허 문서에 대해 한국어 형태소 분석기와 일본어 형태소 분석기를 각각 적용하여 형태소 분석을 수행하고, 사전에 정의된 기술용어 사전을 적용하여 각 특허 문서를 정제한다(②). 이렇게 정제된 문서는 추후 용어의 의미적 유사도 도출을 위해 통합되는데, 양국 특허 문서의 언어가 상이하므로 상호 연결을 위한 전처리를 수행해야 한다. 본 연구에서는 이를 위해 이미 양국에서 동일한 개념을 가리키는 것으로 검증된 기술용어를 지표용어로 활용한다. 구체적으로 정제된 일본 특허 문서에서 검증된 기술용어에 한해 일본어 용어를 한국어 용어로 대체하여 지표화된 일본 특허 문서를 생성한다(③). 이 과정의 필요성



〈Figure 2〉 Overall Process

및 동작 원리는 제 3.3절에서 예를 통해 자세히 소개한다. 다음으로 한국 특허 문서와 지표화된 일본 특허 문서를 통합한 후, Word2Vec 분석을 통해 각 용어를 벡터로 표현한다(④). 마지막으로 각 용어의 벡터간 거리를 산출하여 한국과 일본의 기술 용어 간 유사도 행렬을 도출한다(⑤).

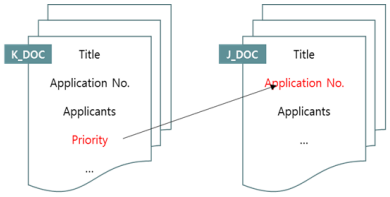
본 장의 이후 절에서는 <Figure 2>의 주요 모듈의 개념 및 동작 원리를 간단한 가상 예를 통해 소개한다.

3.2 우선권 주장 특허 쌍 식별

내용의 유사성이 매우 높은 양국의 특허 문서 쌍을 식별하는 것은 제안 방법론의 매우 중요한 부분 중 하나이다. 본 연구에서는 이를 위해 특허법 중의 우선권 주장 제도를 활용한다. 우선권이란 출원하고자 하는 특허에 대해 출원인이 이미 선출원한 기존 출원건의 출원일로부터 1년 이내에 우선권을 주장하면, 출원하고자 하는 발명을 선출원 시점에 출원한 것으로 판단시점을 인정해주는 특허법상의 권리이다. A 국가에서 선출원한 발명을 B 국가에서 다시 출원하고자 하는 경우 두 출원건은 일반적으로 우선권으로 연결되게 되며, 우선권으로 연결된 양국의 특허는 거의 동일한 내용이 각국의 언어로 기술되어 있다는 특징을 갖는다. 따라서 한국과 일본의 특허 중 우선권으로 연결된 특허는 제안 방법론에 활용되기에 매우 적합한 특징을 갖고 있으며, 본 연구에서는 이들 특허 쌍에 대해 요약 부분을 발췌하여 분석에 사용한다.

본 연구에서 한국과 일본의 특허 쌍을 매핑하기 위해 사용한 구체적 방법은 다음과 같다. 우선 수집의 대상을 일본에서 선출원된 특허를 한국에서 다시 출원하는 발명으로 한정하였다. 이러한 발명은 한국에서 출원될 때 <Figure 3>과 같이 특허의 서지사항에 우선권 정보로서 일본 특허 출원번호가 기재된다.

본 연구에서는 <Figure 3>의 구조를 이용하여 한국과 일본의 특허 쌍을 식별하기 위해 (1) 특정



Title	밀리미터파 감지 센서 (K_DOC)	ミリ波感知センサー (J_DOC)
Application No.	2017-00817	**2009-12345**
Applicants	ABC Corporation	ABC Corporation
Priority	**JP2009-12345**	-

〈Figure 3〉 Mapping with Patent Priority Claim Information

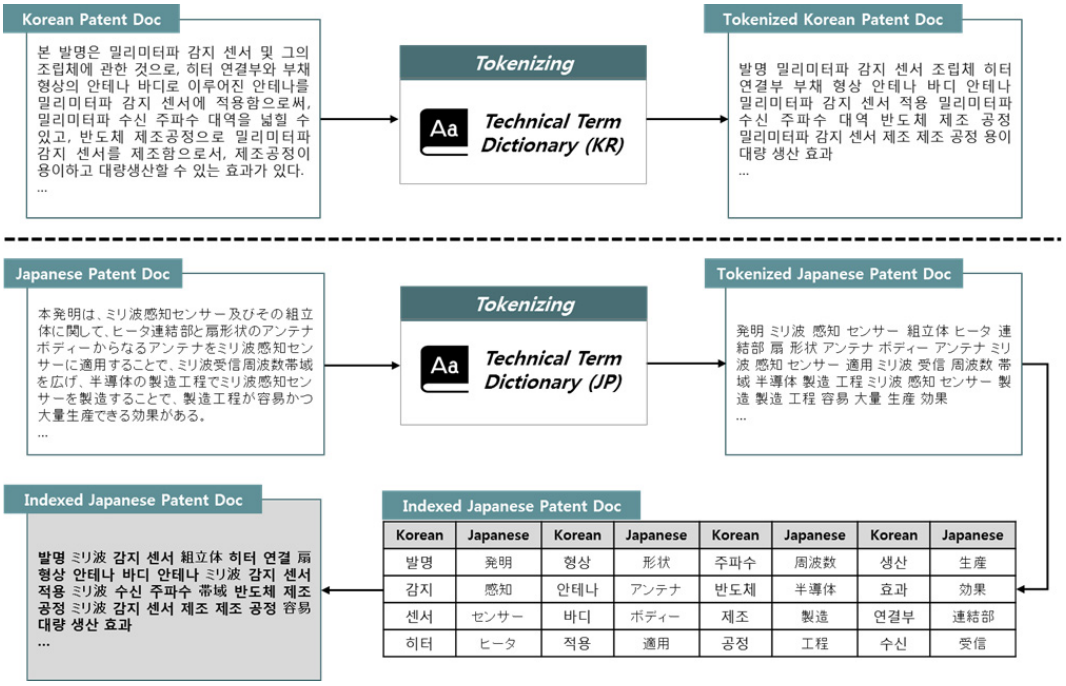
기간 한국에서 출원된 특허 정보를 수집한 후, (2) 우선권 정보에 일본 특허의 출원번호가 기재된 특허만을 추려내고, (3) 해당 일본 특허 출원번호에 대한 검색을 통해 일본 특허 문서를 수집하였다.

3.3 형태소 분석 및 지표용어 삽입

다음은 한일 특허 데이터에 대한 형태소 분석을 거쳐, 특허 문서에서 명사만을 추출하는 과정이다.

형태소 분석은 한국 특허 문서 집합과 일본 특허 문서 집합에 대해 두 차례 이루어진다. 대부분의 기술용어의 품사가 명사인 점에 착안하여, 형태소 분석을 통해 명사 이외의 품사를 갖는 용어를 모두 제거한다. 또한 이들 기술용어 중 관심 용어만을 선정하여 문서를 정제하는데, 관심 용어의 선정은 (1) 빈도수와 (2) 기술용어 여부의 기준에 따라 이루어졌다. 이 과정을 통해 빈도수가 매우 낮거나, 특정 기술과 관련성이 낮은 일반적인 용어들은 모두 제거되었다.

이렇게 도출된 한국 특허와 일본 특허의 관심 용어 리스트는 서로 일대일로 대응되지 않는다. 즉 한국 특허 관심 용어 리스트에만 존재하고 일본 리스트에는 존재하지 않는 용어도 있으며, 한국 특허 관심 용어 리스트에 수록된 하나의 용어가 일본 리스트의 여러 용어에 대응될 수도 있다. 본 연구에서는 위와 같은 상황에서, 한국 특허 관심 용어 리스트의 각 용어에 대응되는 일본 특허 용어를 식별하는 방안을 제시한다. 물론 위의 상황과



〈Figure 4〉 Indexed Japanese Patent Document

반대 방향으로 일본 특허의 관심 용어에 대응되는 한국 특허 용어를 식별하고자 하는 수요도 있을 수 있다. 하지만 이는 제안 방법론의 적용 방향을 반대로 변경함으로써 간단히 수행 가능하므로, 이 과정에 대한 설명은 별도로 다루지 않는다.

이후 과정은 관심 용어만으로 구성된 정제된 특허 문서 쌍을 비교함으로써 용어 간 의미적 유사성을 도출하는 과정이다. 하지만 정제된 문서 집합의 경우 한국어로 작성된 문서와 일본어로 작성된 문서를 모두 포함하고 있으므로, 이들 문서에서 한국어 용어와 일본어 용어의 위치를 상호 비교하기 위한 별도의 매핑 작업이 필요하다. 제안 방법론에서는 이처럼 위치 비교를 위한 매핑에 사용되는 용어를 지표용어라고 명명하였으며, 지표용어는 동일한 기술 개념을 나타내는 양국의 용어 중 이미 널리 알려진 것으로, 서로 일대일로 직역 변환하여도 의미의 손실이 없는 용어로 정의할 수 있다. 즉 지표용어 사전은 한국과 일본 기술용어 간 동의어 사전의 부분집합으로 간주할 수 있다. 제안 방법론은 정제된 일본 특허 문서 중 지표용어 사전에 수록된 일본 기술용어들을 이에 대응되는 한글 기술용어들로 대체하여, <Figure 4>와 같이 지표용어를 삽입한다.

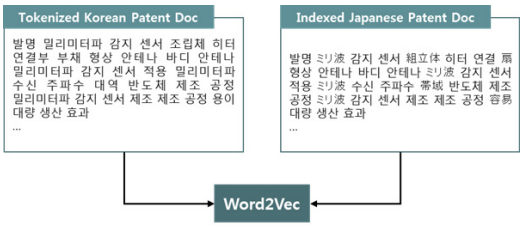
이후 과정에서는 정제된 한국 특허 문서와 지표화된 일본 특허 문서에 포함된 용어의 위치에 기반을 두어 한국과 일본의 용어 간 유사도를 분석하게 된다. 예를 들어 <Figure 4>의 정제된 한국 특허에서 “발명”의 우측에 “밀리미터파”가 위치하고 있으며, 지표화된 일본 특허 문서에서 “發明”의 우측에 “ミリ波”가 위치하고 있다. 이러한 패턴이 빈번하게 발생된다면 “밀리미터파”와 “ミリ波”의 의미적 유사성이 높은 것으로 파악된다.

3.4 Word2Vec 학습을 통한 용어의 벡터화

본 절에서는 Word2Vec 알고리즘을 통해 기술용어를 벡터화하는 과정을 소개한다. Word2Vec 알고리즘에 대한 내용은 이미 많은 연구에서 다루어

졌으므로, 본 연구에서는 상세히 소개하지 않는다. 이전 절에서 소개한 바와 같이 Word2Vec 학습에 사용되는 문서는 <Figure 4>의 “정제된 한국 특허” 문서와 “지표화된 일본 특허 문서”이다. 이들 두 문서 집합을 통합하고 이에 대해 한꺼번에 Word2Vec 학습을 수행함으로써, 양국 특허 문서에 사용된 기술용어를 동일한 관점으로 벡터화할 수 있다. 이러한 과정을 통해 각 용어를 벡터화하고, 각 용어 벡터의 코사인(Cosine) 유사도를 도출한 예가 <Figure 5>에 제시되어 있다.

<Figure 5>의 가운데 표는 각 Word2Vec 알고리즘을 통해 산출된 각 용어의 벡터를 나타내며,



Term	Vector					
감지	0.11	0.01	0.16	0.10	0.05	0.21
주파수	0.04	0.05	0.06	0.04	-0.02	0.15
부채	0.68	-0.07	0.58	0.41	0.03	0.16
扇	0.06	-0.05	0.22	0.06	0.03	0.31
밀리미터파	0.16	-0.06	0.27	0.08	0.04	0.37
ミリ波	0.80	-0.20	0.21	0.17	-0.01	0.76
대역	0.49	-0.50	0.22	0.41	0.39	0.89
帯域	0.50	0.04	0.03	-0.30	0.73	0.14
용이	0.44	0.24	-0.22	0.66	0.41	0.78
容易	0.23	-0.31	0.27	0.41	0.53	0.42
조립체	0.04	0.09	0.50	-0.37	-0.06	0.15
組立体	0.21	-0.57	0.06	-0.09	0.04	0.77

Similarity	부채	밀리미터파	대역	용이	조립체
扇	**0.96**	0.16	0.32	0.07	0.19
ミリ波	0.23	**0.92**	0.24	0.19	0.04
帯域	0.34	0.35	**0.89**	0.26	0.10
容易	0.27	0.28	0.18	**0.95**	0.31
組立体	0.15	0.21	0.09	0.29	**0.94**

〈Figure 5〉 Calculation of Cosine Similarity Using Word2Vec Learning

이 표에는 양국의 기술용어가 혼재되어 있다. 이들 각 용어에 대해 상호간 코사인 유사도를 산출한 결과가 <Figure 5>의 하단 표에 제시되어 있다. 다만 본 연구의 목적이 한국의 각 기술용어와 유사도가 높은 일본의 기술용어를 도출하는 것이므로, 한국 용어 간 또는 일본 용어 간 유사도는 산출할 필요가 없다. 따라서 <Figure 5>의 하단 표는 한국 용어와 일본 용어 간의 유사도만 나타내고 있다. 또한 해당 표에서 각 한국 용어에 대응되는 다섯 개 일본 용어의 유사도에 대해, 유사도가 가장 높게 나타난 일본 용어를 음영 표시하여 나타냈다. 즉 음영 표시된 부분은 각 한국 용어와 유사도가 가장 높게 나타난 일본 용어를 나타내고 있다.

본 장에서는 제안 방법론의 각 요소를 간략한 가상 예를 통해 설명하였으며, 다음 4장에서는 제안 방법론을 실제 데이터를 적용한 실험 과정 및 결과를 소개한다.

4. 실 험

4.1 실험 데이터

본 장에서는 제안 방법론을 실제 특허 데이터에 적용한 실험 과정 및 결과를 소개하며, 본 절에서는 실험에 사용된 데이터를 구체적으로 소개한다.

우선 분석의 기준이 되는 한국 특허 문서는 특허정보검색서비스를 제공하는 KIPRIS 시스템에서 검색을 통해 수집하였다. 수집 시 특허의 출원일이나 출원인 등에는 별도의 제약을 두지 않았으며, 발명이 속한 기술 분야에 대해서도 제한을 두지 않고 전체 분야에 대한 검색을 수행하였다. 다만 본 실험은 특허 명세서의 요약 부분을 수집하여 활용하기 때문에, 요약의 대부분이 화학식 등의 기호로 구성되어 본 실험에 적합하지 않다고 판단한 데이터는 실험에서 제외하였다.

또한 제안 방법론은 동일한 발명이 양국의 언어로 출원된 데이터에 대한 분석을 필요로 하므로,

수집한 데이터 중 선출원된 일본 특허에 대해 우선권을 명시적으로 주장한 특허만을 발췌하였다. 또한 이들 특허의 서지사항에 기재된 우선권 정보를 통해 일본특허검색사이트인 특허정보플랫폼 J-PlatPat과 WIPO로부터 선출원된 일본 특허를 수집하고 요약 부분을 발췌하였다.

이상의 과정을 통해 최종적으로 실제 실험에는 우선권으로 매칭된 한국 특허 문서 343건, 일본 특허 데이터 343건, 총 686건의 특허 문서를 사용하였다. 수집한 데이터 원본 내용의 일부를 <Figure 6>에 제시하였다.

Korean Patent Documents (Sample)

DOCNO	CONTENTS
1	게이트웨이, 다수의 기지국들, 및 모바일 디바이스를 포함하는 통신
2	본 발명이 해결하고자 하는 과제는 복수의 파스너에 하나씩 RFID 태
3	무선 기지국은 공통 채널을 이용하여 복수의 이동국에 동일한 데이
4	본 발명의 목적은 다양화된 정보 전달 요구에 대응하기 위해서, 다
5	비디오 인코더는 입력 동영상의 영상 데이터를 예측에 기초하여 인
6	판정부는, 통신 어플리케이션의 취득에 이용한 제 네트워크의 정보
7	반도체 패키지가 실장된 기판을 곡면화해도, 전기적 접속부에 대한
8	모바일 통신 디바이스가 다른 모바일 통신 디바이스와의 디바이스
9	제 무선국은, 제 셀에서의 무선 단말과의 제 무선 접속이 확립되어
10	코어 네트워크 내에 배치된 네트워크 노드는 트리거 메시지를 MTC

Japanese Patent Documents (Sample)

DOCNO	CONTENTS
1	ゲートウェイと、複数の基地局と、移動デバイスとを備える通信
2	本発明が解決しようとする課題は、上記問題点を解決することで
3	無線基地局は、共通チャネルを利用して複数の移動局に同一のデ
4	多様化された情報配信へのニーズに対応するために、様々な事業
5	映像符号化器は、入力された動画像における画像データを予測に
6	判定部は、通信アプリケーションの取得に用いた第ネットワーク
7	半導体パッケージが実装された基板を曲面化しても、電気的接続
8	移動通信デバイスが別の移動通信デバイスとデバイス間ＤＤ通信
9	第の無線局は、第のセルでの無線端末との第の無線接続が確立され
10	コアネットワーク内に設置されるネットワークノードは、トリガ

〈Figure 6〉 Korean/Japanese Patent Documents(Part)

4.2 형태소 분석을 통한 기술용어 추출 및 관심 용어 리스트 작성

본 절에서는 <Figure 6>의 686건의 특허에 대한 형태소 분석을 통해 한국/일본의 기술용어를 추출하고 관심 용어 리스트를 도출한 결과를 소개한다. 형태소 분석은 양국의 문서에 대해 각각 수행되었으며, 구체적으로 한국 특허 문서에 대한 형태소 분석은 Komoran 형태소 분석기를, 일본 특허 문서에 대한 형태소 분석은 오픈 소스인 Mecab 형태소 분석기를 사용하여 명사를 추출하였다.

1차로 추출된 명사 중 빈도수가 높고 기술용어임이 명확한 용어만을 추려서 관심 용어 리스트를 작성하였다. 구체적으로는 우선 각 명사의 빈도수를 확인하여, 빈도수 9 이하인 명사는 모두 분석에서 제외하였다. 이후 빈도수 10 이상인 명사 중에서 특허 기술용어라기 보다는 일상적으로 많이 사용되는 일반 용어에 가까운 용어들은 수작업으로 식별하여 제거하였다.

위의 과정을 통해 한국의 관심 용어 72개를 선별하고, 이에 대응되는 일본의 용어를 식별하였다. 이들 관심 용어 리스트는 본 실험에서 두 가지 용도로 사용된다. 첫 번째 용도는 관심 용어 고유의 목적, 즉 한국 기술용어에 대응되는 일본 기술용어를 얼마나 잘 식별하는지 판단하기 위한 목적으로 활용되는 것이다. 또 다른 용도로 관심 용어 리스트는 지표용어 삽입을 위한 양국 간 기술용어의 동의어 사전으로도 활용될 수 있다.

Indexed Japanese Patent Doc

| | | |

(18개 관심 용어) *계수, 냉각, 데이터, 무선, 변수, 변환, 부하, 송출, 식별자, 앵커, 원격, 저장(記憶, 保持, 格納), 전송(伝送, 送信), 전압, 증폭, 통신, 할당(割り当て, 割当), 행렬*

위에서 저장, 전송, 할당 등 하나의 한국 기술용어가 여러 개의 일본 기술용어에 대응되는 경우, 괄호 내에 대응 가능한 일본 기술용어를 모두 기재하였다.

4.3 특허 문서의 정제 및 지표용어 삽입

본 절에서는 앞에서 생성한 관심 용어 리스트 및 지표용어를 활용하여 특허 문서를 정제하고 지표화된 일본 특허 문서를 생성한 결과를 소개한다.

우선 <Figure 7>은 72개 쌍으로 구성된 관심 용어 리스트를 적용하여 <Figure 6>의 특허 원문을 정제한 결과를 보이고 있다.

Tokenized Korean Patent Doc

게이트웨이 기지국 디바이스 통신 게이트웨이 기지국 데이터 기지국 식별 기지국 기지
무선 기지국 채널 데이터 송신 무선 기지국 무선 기지국 송신 데이터 송신 데이터 송신
통신 네트워크 통신 통신 통신 네트워크 네트워크 통신 네트워크 통신 네트워크 통신
반도체 기판 접속 접속 접속 반도체 전극 반도체 반도체 반도체 반도체 전극 접속
무선국 무선 단말 무선 접속 무선 단말 통신 무선국 무선국 무선 단말 무선 접속 무선

Tokenized Japanese Patent Doc

ゲートウェイ 基地局 デバイス 通信 ゲートウェイ 基地局 データ 基地局 識別 保持 基
無線 基地局 チャネル データ 送信 無線 基地局 無線 基地局 送信 データ 送信 データ
通信 ネットワーク 通信 通信 通信 ネットワーク ネットワーク 通信 ネットワーク 通
半導体 基板 接続 接続 接続 半導体 電極 半導体 半導体 半導体 半導体 電極 接続
無線局 無線 端末 無線 接続 無線 端末 通信 無線局 無線局 無線 端末 無線 接続 無線

〈Figure 7〉 Tokenized Patent Documents(Part)

다음으로 <Figure 7>의 정제된 일본 특허 문서에 사용된 일본 기술용어 중 54개의 지표용어에 포함된 용어들을 각각에 대응되는 한국 기술용어로 대체한 결과의 일부가 <Figure 8>에 나타나 있다.

<Figure 8>에서 지표용어에 해당되는 용어는 한국어로, 지표용어에 해당되지 않은 용어는 일본어로 표시되어 있다. 즉 지표화된 일본 특허 문서는 한국어와 일본어가 혼재된 형태로 표현되어 있다.

4.4 용어 벡터화 및 용어 간 유사도 도출

본 절에서는 Word2Vec 학습을 통해 각 기술용어를 벡터화하고, 최종적으로 한국 관심 용어 18개에 대응하는 일본 관심 용어의 유사도를 산출한 실험 결과를 소개한다. 전술한 바와 같이 Word2Vec 학습은 <Figure 7> 상단의 정제된 한국 특허 문서와 <Figure 8>의 지표화된 일본 특허 문서를 통합한 집합에 대해 수행된다.

Word2Vec 학습은 Python 3.6 버전에서 Gensim 라이브러리를 사용하여 구현하였다. 또한 용어 출현의 최소 빈도수로는 1을 적용하였으며, 각 용어를 300차원의 벡터스페이스에 임베딩하였다. 또한 Word2Vec 알고리즘을 구현한 CBOW와 Skip-gram의 두 가지 학습 모델 중, 본 실험에서는 Skip-gram 모델을 사용하여 용어를 벡터화하였다.

<Figure 9>는 학습을 통해 벡터화한 용어의 일부 형태를 나타낸다.

Term	Vector1	Vector2	Vector3	Vector4	Vector5	...
게이트웨이	-0.20971	0.02960	0.07392	-0.17710	0.00081	
기지국	-0.12376	0.03513	0.21413	-0.00538	-0.03477	
디바이스	-0.33623	0.05422	-0.09758	-0.19855	0.00731	
통신	-0.36986	-0.05201	0.00448	-0.17158	-0.07304	
데이터	-0.17211	-0.01000	0.01239	-0.23224	-0.05037	
通信	-0.16745	0.10163	-0.00322	-0.09169	0.04925	...
データ	0.03136	0.10071	-0.02165	-0.07220	0.04178	
保持	-0.12782	0.00834	0.02569	-0.09821	-0.00699	
送信	-0.01994	0.07814	0.04289	0.05028	-0.04114	
無線	0.08809	0.02367	0.14228	0.05290	0.04395	
...			...			

〈Figure 9〉 Term Vectors(Part)

다음으로 이들 용어 벡터 간 코사인 유사도를 산출하였다. 우선 벡터화된 용어 중 지표용어, 즉 이미 양국간 동의어가 식별된 용어는 유사도 산출 대상에서 제외하였다. 또한 전술한 바와 같이 한국 용어 간 또는 일본 용어 간의 유사도도 산출 대상에서 제외하였다. 즉 지표용어에 포함되지 않은 용어들에 대해서 한국/일본 용어 간 유사도만을 산출하게 되었으며, 그 결과의 일부를 <Figure 10>에 제시하였다. 유사도 값은 0에서 1 사이의 값을 가지며, 1에 가까울수록 해당 용어 간 유사도가 높은 것으로 해석된다.

	계수	냉각	데이터	무선	변수	변환	...
冷却	0.5023	0.7801	0.3889	0.4465	0.7207	0.4706	
遠隔	0.0282	0.5269	0.3167	0.6714	0.2040	0.0493	
割当	0.5687	0.8168	0.3742	0.4916	0.5294	0.4495	
割り当て	0.4878	0.7049	0.1724	0.3933	0.4312	0.3223	
負荷	0.0305	0.3986	0.1877	0.3810	0.2541	0.0626	
識別子	0.2229	0.3168	0.4559	0.1013	0.4242	0.3938	
アンカー	0.0739	0.5165	0.2870	0.6574	0.1030	0.0498	
通信	0.0520	0.3809	0.1346	0.2880	0.0885	0.0071	
保持	0.1219	0.4031	0.1958	0.2735	0.2488	0.1368	
変数	0.6578	0.7113	0.2392	0.0344	0.8339	0.5243	
電圧	0.5248	0.7348	0.1541	0.2576	0.6540	0.3786	
格納	0.1373	0.3227	0.2026	0.1569	0.3121	0.1813	...
無線	0.0687	0.5327	0.1000	0.5145	0.0491	-0.0692	
送出	0.3627	0.6234	0.1307	0.3970	0.2952	0.1972	
記憶	0.2465	0.3097	0.2405	0.0147	0.3672	0.3195	
データ	0.2311	0.3968	0.2710	0.1412	0.3045	0.2599	
増幅	0.4928	0.7029	0.0424	0.2943	0.4854	0.2882	
係数	0.7817	0.7214	0.1989	0.1024	0.6687	0.5471	
変換	0.5802	0.5030	0.2509	-0.0639	0.5709	0.5440	
伝送	0.3084	0.4686	0.1647	0.1982	0.2503	0.1856	
送信	0.2662	0.4798	0.0227	0.2123	0.1654	0.0924	
行列	0.6905	0.7267	0.1870	0.0674	0.5562	0.5115	

〈Figure 10〉 Similarity between Korean/Japanese Terms(Part)

4.5 실험 결과 분석

본 절에서는 제안 방법론의 성능을 파악하기 위해 한국어 관심 용어 18개에 대응되는 일본어 용어 22개의 유사도를 분석하였다. 한국어 용어 중 “저장”, “전송”, 그리고 “할당”의 경우 각각 복수개의 일본어 용어에 대응되므로, 사용된 의미에 따라 서로 다른 용어로 간주하였다. 즉 22개의 한국어 용어에 대응되는 일본어 용어를 얼마나 잘 식별하는지 분석함으로써, 제안 방법론의 성능을 평가하고자 한다. <Figure 11>은 22개 한국어 용어에 대응되는 일본어 용어의 식별 결과를 보이고 있다.

<Figure 11>에서 빈도수는 해당 한국어 용어가 출현한 횟수, 순위는 전체 22개 용어 가운데 해당 한국어 용어에 가장 잘 대응되는 것으로 전문가가 선별한 일본어 용어의 순위, 그리고 상위(%)는 위의 일본어 용어의 순위를 전체 용어 수인 22개로 나눈 비율을 뜻한다. 예를 들어 한국어 용어 “앵커”의 경우 전문가는 이에 대응되는 일본어 용어로 “アンカー”를 선정하였다. “앵커”라는 용어는 총 10회 출현하였으며, 일본어 용어 22개 각각이 “앵커”와 갖는 유사도 가운데 “アンカー”와 “앵커”

Term	계수	냉각	데이터	무선	변수
Frequency	39	10	241	326	16
Rank	1	2	6	3	1
High Rank(%)	5%	9%	27%	14%	5%

Term	변환	부하	송출	식별자	앵커	원격
Frequency	64	28	15	45	10	18
Rank	2	2	1	1	1	2
High Rank(%)	9%	9%	5%	5%	5%	9%

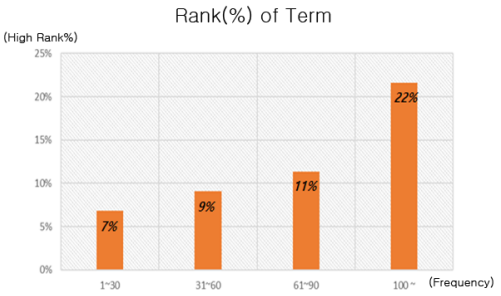
Term	저장 (記憶)	저장 (保持)	저장 (格納)	전송 (伝送)	전송 (送信)
Frequency	32	22	25	65	296
Rank	4	3	1	3	8
High Rank(%)	18%	14%	5%	14%	36%

Term	전압	증폭	통신	할당 (割り当て)	할당 (割当)	행렬
Frequency	18	19	410	16	6	20
Rank	1	1	2	1	2	1
High Rank(%)	5%	5%	9%	5%	9%	5%

〈Figure 11〉 Rank of Corresponding Terms of the Core Terms

의 유사도가 가장 높은 순위인 1위를 차지하였음을 알 수 있다. 유사도 1위는 전체 22개 용어 중 상위 5% 이내에 해당된다.

<Figure 11>에서 나타난 바와 같이 일부 용어는 대응 용어를 유사도 순위 1위로 정확하게 식별하였으나, 일부 다른 용어의 경우 대응 용어의 유사도 순위가 6위, 8위 등으로 낮게 나타난 경우도 확인하였다. 따라서 이러한 정확도의 양상을 살펴보기 위해, 각 용어의 출현 빈도수에 따른 대응 용어의 순위 비율을 분석하여 <Figure 12>에 제시하였다.



〈Figure 12〉 Rank(%) of Term Distribution

<Figure 12>에서 각 용어를 출현 빈도에 따라 네 그룹으로 묶어서 가로축에 표시하였다. 또한 <Figure 12>의 세로축은 각 그룹의 용어들에 대해 대응 용어

의 순위 평균을 표시한 것이다. 예를 들어 “30~60” 구간에는 “계수”, “식별자”, 그리고 “저장(記憶)”의 세 개의 용어가 포함되어 있음을 <Figure 11>에서 확인할 수 있으며, 이들의 대응 용어의 순위 상위(%)는 각각 5%, 5%, 그리고 18%임을 알 수 있다. 따라서 <Figure 12>에서 이들의 순위 평균은 상위 9%로 나타난다.

<Figure 12>에서 출현 빈도가 낮은 용어일수록 대응 용어의 순위(%)가 나타나므로, 의미적 유사성에 기반을 둔 유사어 식별이 정확하게 이루어지는 것으로 이해할 수 있다. 반면 출현 빈도가 높은 용어의 경우 대응 용어의 식별 정확도가 상대적으로 낮게 나타남을 확인하였다. 이러한 현상이 발생하는 이유는 출현 빈도가 높은 용어일수록 이미 충분히 보편화된 기술용어일 가능성이 크며, 따라서 각 국가에서 이들 용어를 고유의 기술 개념뿐 아니라 확장된 의미로 활용할 수 있다는 측면에서 찾아볼 수 있을 것이다 .

5. 결 론

사전에 기반을 둔 정적 다국어 검색 방식은 동일한 기술 개념이 국가별로 상이한 어휘로 표현될 때 관련 선행기술을 충분하게 찾아내지 못한다는 한계를 보인다. 이러한 현상은 새로운 기술 개념이 빠르게 출현하며, 이들 개념에 대한 용어의 표준화가 충분히 이루어지기도 전에 다수의 특허가 출원되는 최근 환경에서 더욱 큰 한계로 나타날 수 있다.

본 연구에서는 특허 출원 전후에 이루어지는 선행기술조사 시, 기존의 정적 사전 기반 다국어 검색 방식의 한계를 극복하기 위한 방안을 제시하였다. 구체적으로 한국과 일본의 실제 특허 문서에 사용된 용어의 용례 분석을 통해 한국어와 일본어로 기술된 기술용어의 유사성을 산출하고, 이를 기반으로 의미 기반 해외 유사 특허를 검색할 수 있는 방안을 제시하였다.

학술적 측면에서 본 연구는 두 개의 상이한 언어체계를 갖는 문서로부터 용어의 의미적 유사도를 산출하는 방안을 제시했다는 점에서 기여를 인정받을 수 있다. 가장 큰 학술적 기여는 우선권 주장으로 연결된 특허 문서 쌍을 분석에 사용하였다는 점과, 지표용어 삽입이라는 새로운 개념을 도입하여 한국어와 일본어로 각각 기술된 문서의 통합 분석을 가능하게 하였다는 점에서 찾을 수 있다. 또한 특허 데이터를 활용하여 해당 분야인 특허 분야 프로세스의 개선에 활용할 수 있는 방법론을 제안하였다는 점과 상이한 언어로부터 유사도를 도출하기 위해 각각의 형태소 분석기를 별도로 사용하여 데이터를 정제한 후 이를 통합하는 방안 제안하였다는 점도 학술적 기여로 인정받을 수 있다.

본 연구의 기여는 실무적 측면에서 더욱 인정받을 수 있다. 해외 유사 특허 검색 시 특허분류코드에 의한 검색과 키워드 및 검색식을 사용한 검색이 병행 사용되지만, 이러한 방식은 전술한 바와 같은 측면에서 여러 한계를 갖는다. 향후 동일 개념이 양국에서 다른 용어로 표현되는 경우의 검색, 그리고 양국 간 동의어 사전에 아직 수록되지 않은 최신 용어의 검색 등이 필요한 상황에서 본 연구에서 제안한 방법론이 특허 선행기술조사 검색의 업무 효율성을 높여줄 것으로 기대한다.

하지만 제안 방법론은 다음과 같은 측면에서 향후 보완이 필요하다. 우선 Word2Vec 학습은 본질적으로 대량의 학습 데이터를 필요로 한다. 본 연구에서는 실제 출원된 특허 문서 가운데 우선권으로 연결된 총 686개만의 문서를 실험에 활용하였다. 따라서 제안 방법론의 견고성과 정확성을 향상시키기 위해서는 훨씬 더 방대한 양의 문서에 대한 반복 실험이 반드시 수행되어야 한다. 또한 제안 방법론은 한국어와 일본어의 경우와 같이 어순이 동일한 언어체계를 갖는 문서 집합에만 적용될 수 있으며, 양국의 언어로 동일한 내용을 다루고 있는 문서에 기반을 두고 동작한다는 특징을 갖는다. 또한 본 연구에서는 제안 방법론의 동작 원리 및 적용 사례만을 제시하였으나, 향후 유사

선행 기술과의 성능 비교를 통해 제안 방법론의 우수성을 보일 필요가 있다. 또한 향후 특허 분야의 우선권 주장 외에도, 이러한 특징을 갖고 있는 다양한 문서 집합을 발굴하여 분석의 수월성을 향상시킬 필요가 있다.

References

Baik, J.B., S.M. Kim, and S.W. Lee, "Extracting Alternative Word Candidates for Patent Information Search", *Journal of KIISE : Computing Practices and Letters,* Vol.15, No.4, 2009, 299-303.

(백종범, 김성민, 이수원, "특허 정보검색을 위한 대체어 후보 추출 방법", *정보과학회논문지 : 컴퓨팅의 실제 및 레터,* 제15권, 제4호, 2009, 299-303.)

Bergmann, I., D. Butzke, L. Walter, J.P. Fuerste, M.G. Moehrle, and V.A. Erdmann, "Evaluating the Risk of Patent Infringement by Means of Semantic Patent Analysis : The Case of DNA Chips", *R&D Management,* Vol.38, No.5, 2008, 550-562.

Chang, J.K., "Study on the Management of Utility Model System", *The Journal of Intellectual Property,* Vol.12, No.4, 2017, 103-134.

(장진규, "4차 산업혁명 시대의 실용실안제도 운용방안", *지식재산연구지,* 제12권, 제4호, 2017, 103-134.)

Jun, S.H., "Technology Forecasting of Intelligent Systems Using Patent Analysis", *Journal of Korean Institute of Intelligent Systems,* Vol.21, No.1, 2011, 100-105.

(전성해, "특허분석을 이용한 지능형시스템의 기술예측", *한국지능시스템학회논문지,* 제21권, 제1호, 2011, 100-105.)

Kim, H.J., "Technology Trends Related with the Natural Dyeing by Patent Analysis", *Fashion & Textile Research Journal,* Vol.11, No.6, 2009, 942-946.

(김호정, "연구논문 : 특허분석을 통한 천연염색기술의 현황", *한국의류산업학회지,* 제11권, 제6호, 2009, 942-946.)

Kim, P.R. and S.H. Hwang, "A Study on the Projection of the IT-based Promising Technologies Utilizing Patent Database", *The Journal of Korean Institute of Communications and Information Sciences,* Vol.34, No.10, 2009, 1021-1030.

(김방룡, 황성현, "특허 정보를 활용한 IT 유망기술 도출에 관한 연구", *한국통신학회논문지,* 제34권, 제10호, 2009, 1021-1030.)

Lee, D.H., H.K. Boo, and N.G. Kim, "Predicting Missing Ratings of Each Evaluation Criteria for Hotel by Analyzing User Reviews", *Journal of Information Technology Services,* Vol.16, No.4, 2017, 161-176.

(이동훈, 부현경, 김남규, "사용자 리뷰 분석을 통한 호텔 평가 항목별 누락 평점 예측 방법론", *한국IT서비스학회*, 제16권, 제4호, 2017, 161-176.)

Lee, K.B., S.G. Jeon, N.S. Nho, K.H. Kim, D.H. Shin, S.W. Kim, and Y.H. Kim, "Patent Analysis of Oil Sands Bitumen Upgrading Technologies", *Journal of the Korean Industrial and Engineering Chemistry,* Vol.19, No.6, 2008, 592-599.

(이기봉, 전상구, 노남선, 김광호, 신대현, 김선욱, 김용헌, "오일샌드 역청 개질 기술의 특허정보 분석", *Journal of the Korean Industrial and Engineering Chemistry,* 제19권, 제6호, 2008, 592-599.)

Mikolov, T., K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space", arXiv preprint arXiv, 2013.

Moon, J.S., M.H. Cha, Y.S. Yoon, and S.M. Choi,

"Patent and Paper Information Analysis : Questionnaire and Program for Syndrome Differentiation", *Korea Journal of Oriental Medicine,* Vol.12, No.3, 2006, 17-29.

(문진석, 차민호, 윤유식, 최선미, "辨證을 위한 설문지 및 프로그램에 대한 특허정보 및 논문을 통한 연구 동향 분석", *한국한의학연구원논문집,* 제12권, 제3호, 2006, 17-29.)

Park, C.C. and H.J. Kim, "Development Trends of Domestic Apparel Design by Analyzing Patent Applications", *Fashion and Textile Research Journal,* Vol.12, No.4, 2010, 508-512.

(박차철, 김호정, "연구논문 : 특허정보 분석을 통한 국내 의류 디자인 개발 동향", *한국의류산업학회지,* 제12권, 제4호, 2010, 508-512.)

Park, C.C. and H.J. Kim, "Development Trends of Korean Textile Industry by Analyzing Domestic Patent", *Fashion and Textile Research Journal,* Vol.11, No.5, 2009, 840-845.

(박차철, 김호정, "연구논문 : 특허분석을 통한 한국 섬유산업 기술개발 동향", *한국의류산업학회지,* 제11권, 제5호, 2009, 840-845.)

Park, H.W. and K.I. Kim, "Analysis of Research Trends and Technological Position of PMP Using Patent Information", *The Journal of the Korea Contents Association,* Vol.7, No.9, 2007, 117-126.

(박현우, 김기일, "특허정보를 통한 PMP 연구동향과 기술경쟁력 분석", *한국콘텐츠학회논문지,* 제7권, 제9호, 2007, 117-126.)

Park, H.W. and O.H. Sung, "An Analysis of Patterns of Knowledge Flows of Science and Technology Using Scientific Papers and Patents", *Knowledge Management Research,* Vol.9, No.1, 2008, 39-59.

(박현우, 성웅현, "과학논문과 특허를 통한 과학기술 지식흐름의 특성분석", *지식경영연구,* 제9권, 제1호, 2008, 39-59.)

Wong, W.X.S. and N.G. Kim, "R&D Perspective Social Issue Packaging using Text Analysis", *Journal of Information Technology Services,* Vol.15, No.3, 2016, 77-95.

Yoon, B.S., D.R. Lee, and B.M. Yun, "The Effect of the Patent Acquisition on the Management Performance", *The Journal of Business Education,* Vol.24, No.4, 2010, 261-288.

(윤병섭, 이덕록, 윤복만, "특허취득 활동이 경영성과에 미치는 효과 분석", *상업교육연구,* 제24권, 제4호, 2010, 261-288.)

Zhang, J.L., J.H. Jang, S.J. Kim, H.K. Lee, and C.H. Lee, "A Study on the Efficient Patent Search Process Using Big Data Analysis Tool R", *Journal of Korea Safety Management and Science,* Vol.15, No.4, 2013, 289-294.

(장청윤, 장정환, 김석주, 이현근, 이창호, "빅데이터 분석 도구 R을 활용한 효율적인 특허 검색에 관한 연구", *대한안전경영과학회지,* 제15권, 제4호, 2013, 289-294.)

◈ About the Authors ◈

Minji Paek (pmj7320@kookmin.ac.kr)

Minji Paek received the B.S. degree in Japanese from Semyung University in 2010 and Master degree in Business IT from Kookmin University in 2018. Her current research interests include text mining, data mining and patent analysis.

Namgyu Kim (ngkim@kookmin.ac.kr)

Professor Namgyu Kim received the B.S. degree in Computer Engineering from Seoul National University in 1998 and Ph.D. degree in Management Engineering from Korea Advanced Institute of Science and Technology (KAIST) in 2007. He has been working for Kookmin University since then. His current research interests include text mining, data mining, and data modeling.