선행기술 검색을 위한 특허 명세 키워드의 추출

이지석, 김경재, 박소현, 박영연, 이우석, 임현우, 정남주, 성영락, 박준석
국민대학교

runseok23@kookmin.ac.kr, economy02@kookmin.ac.kr, sohyunzzq@kookmin.ac.kr,
parkyoungyeo@kookmin.ac.kr, wolfman352@kookmin.ac.kr, dla418@kookmin.ac.kr,
plmnko1016@kookmin.ac.kr, yeong@kookmin.ac.kr, jspark@kookmin.ac.kr

Extracting Keywords from Patent Specifications for Prior Art Searching

Lee Ji Seok, Kim Gyung Jae, Park So Hyun, Park Young Yeon, Lee Woo Seok,
Lim Hyun Woo, Jung Nam Joo, Seong Yeong Rak, Park Jun Seok
Kookmin Univ.

요 약

본 논문은 특허 명세서에서 필요로 하는 선행 기술 문헌을 작성하는데 있어 여러 가지 문제들을 해결하기 위해 TF-IDF(Term Frequency-Inverse Document Frequency) 키워드 추출 방식과 KIPRIS(Korea Intellectual Property Rights Information Service) Plus-Open API 를 활용하는 방법을 제안한다. 먼저 명칭을 활용한 키워드 추출 방식에 TF-IDF 라는 방식이 적합한지 확인하고, 키워드 추출 시 필요한 데이터 추출 및 전처리 과정을 통해 해당 방식들의 성능을 높이고자 하였다. 그 후 KIPRIS Plus-Open API 를 이용하여 관련된 선행 기술을 탐색함으로써 실제 특허 명세서에 작성된 선행기술조사 문헌과 유사하게 나타나는 것을 확인했다. 그 결과 기존에 변리사가 직접 구하는 방식에 비해 시간과 효율을 높여준다.

Ⅰ. 서 론

특허 명세서를 작성하는데 있어서 선행기술조사 문헌을 작성하는 것이 권장된다. 이는 출원에 선행하는 같은 내용의 기술 또는 선출원이 존재하는 지의 여부에 대해 특허 문헌을 정밀히 조사하여 특허성을 파악하는 데 있어 필수요소이다. 문서를 조사하는 과정 및 해당 문서들이 선행기술 문헌으로서 적합한지 확인하는 것에는 많은 시간이 소요된다. 본 논문에서는 이러한 문제점을 해결하고자 특허 명세서 데이터를 전처리하고 발명의 명칭을 이용해 키워드 추출 방식 중 하나인 TF-IDF 방식을 활용하여 선행기술조사 문헌을 작성할 때 필요한 키워드를 추출하고자 한다. 추출된 키워드를 KIPRIS Plus-Open API 에 활용하여 검색하였다. 이를 통해 검색 시간을 절약하고, 정보 탐색의 한계를 극복하고자 하였다. 이에 본 논문에서는 우선 특허 명세서 데이터의 전처리 과정을 통해 발명의 명칭을 명사화 하는 과정을 설명한다. 전처리된 데이터에서 TF-IDF 키워드 추출 방식을 활용하여 키워드를 추출한 결과를 보여준다. 결론적으로 추출한 키워드를 KIPRIS Plus-Open API 를 활용하여 작성된 선행기술조사 문헌을 실제 특허 명세서와 비교하여 타당성을 검증하고자 한다.

Ⅱ. 본 론

본 논문은 TF-IDF 키워드 방식을 활용하여 선행기술조사 문헌을 작성하고자 우선 발명 신고서의 데이터를 전처리하는 과정을 거쳤다. 띄어쓰기와 단어의 변화가 적은 영어와 달리, 한국어는 문법적으로 복잡하다. 따라서 한국어 문장의 형태소를 분석하기 위해서는 한국어 문법에 대해 깊이 있는 이해력이 필요하다. 이를 위해 한국어 자연어 처리를 위한 파이썬 패키지인 KoNLPy[1]를 사용해 데이터 전처리를 수행하였다. 해당 패키지는 꼬꼬마, 한나눔, Komoran, MeCap-ko, OKT를 활용하며, 형태소 분석, 품사 태깅 기능 또한 존재한다. 본 논문에서 실시한 데이터 전처리를 위해서 KoNLPy에 내장된 다른 형태소 처리기들보다 형태소 처리 속도가 빠르고 형태소 분석 시 태깅하는 품사의 범위가 포괄적인 높은 OKT(Twitter) 클래스를 사용해서 데이터의 품사화를 구현하고자 했다. 하여, 약 303개의 특허문서 데이터 셋을 KIPRIS와 인터넷 크롤링을 통해 CSV 파일로 구성하였고, KoNLPy의 OKT 클래스를 활용해 데이터를 전처리하였다.

전처리 전: 무용제 폴리우레탄 합성피혁과 그 제조방법
전처리 후: 무용, 제, 폴리우레탄, 합성, 피혁, 제조방법

그림 1. KoNLPy의 OKT 클래스를 활용하여 출력한 발명의 명칭 데이터

그림 1에서 “무용제 폴리우레탄 합성피혁과 그 제조방법”이라고 입력했을 때, ‘과’ 혹은 ‘그’ 와 같은 불용어를 제거하고, 각 단어를 명사화 하여 출력하는 것을 볼 수 있다.

앞서 실시한 방식으로 전처리된 데이터를 TF-IDF 방식

을 활용하여 키워드를 추출하였다. 여기서 TF-IDF[2]는 특정 단어의 중요도를 계산하는 데 사용되는 통계적 척도로 단어의 출현 빈도인 TF(Term Frequency)와 문서 내 전체 단어 수의 역수인 IDF(Inverse Document Frequency)를 곱하여 계산한다. TF-IDF를 활용해 불용어가 제거된 특허문서 데이터 셋에서 발명의 명칭을 이용하여 각 특허문서 데이터의 TF-IDF 값을 분석했다. 결과적으로 TF-IDF의 값이 클수록 문장 안에서 주요한 단어일 가능성이 크다는 것을 확인할 수 있다.

Document 76: 스웨드- [TF-IDF]:0.6172576006861191 폴리에스터- [TF-IDF]:0.617257600686119 제조방법- [TF-IDF]:0.20501086560103277 인조- [TF-IDF]:0.4426699151466545
(a) 문서 76의 특허 명세서 제목을 통한 TF-IDF
76,20010205,1999,폴리에스터 인조 스웨드의 제조방법 77,,1999,모피형 합성피혁,D,,20000306,D06N,D06N-003/ 78,,1999,입체무늬가 돌출형성된 가죽소재,C,,20000316, 79,19991105,1999,인공피혁 제조방법,D,1019990078892, 80,20010315,1999,통기성이 있는 인조피혁과 그 제조방법
(b) 특허 명세서 데이터 셋

그림 2. 특허명세서의 TF-IDF 진행 전후

그림 2(a)은 ‘폴리에스터 인조 스웨드의 제조방법’ 이라는 발명의 명칭을 TF-IDF 방식을 사용하여 분석한 결과이다. 여기서 키워드는 TF-IDF 값이 크게 나온 ‘스웨드’ 와 ‘폴리에스터’를 꼽을 수 있다. 그에 반하여 ‘제조방법’ 은 그림 2(b)에서와 같이 특허 명세서의 발명의 명칭에서 자주 등장하는 단어로 TF-IDF 값이 작게 나오는 것을 확인할 수 있다. 따라서 TF-IDF 방식은 분야에 따라 세세하게 분류되어 있는 특허 명세서의 특징을 활용하여, 각 특허 명세서의 발명의 명칭에서 불용어를 제거하고 명사화 된 단어의 집합들을 사용하여 한 문서에서 많이 출현하고, 전체 문서에서 적게 출현하는 단어를 키워드로 추출해낼 수 있다. 그리하여 TF-IDF가 발명의 명칭을 활용해 키워드를 추출할 때 적합한 방식임을 알 수 있다.

TF-IDF 방식을 활용해 추출된 키워드는 공백으로 분리되어 있는데, KIPRIS Plus-Open API에 검색식으로 사용하기 위해서 공백을 ‘*’로 바꾸는 후처리 과정을 거쳤다. 하여, KIPRIS Plus-Open API에서 선행기술 문헌을 검색할 때 모든 키워드가 포함되도록 검색식을 만들었다. 검색 결과에서 출력된 유사 특허 명세서의 발명의 명칭과 출원번호만 후처리를 통해 저장했다.

Prior Art (특허문헌 0001) 한국등록특허 제10-1996-0020264호 <농색성 및 견뢰도가 우수한 스웨드조 폴리에스터 직물의제조방법> (특허문헌 0002) 한국등록특허 제10-2005-0122950호 <나일론해도사 및 일반폴리에스터사를 사용한 스웨드조인공피혁의 제조방법>
(c) ‘폴리에스터 인조 스웨드의 제조방법’을 넣고 선행 기술 문헌을 탐색한 결과

그림 3. KIPRIS API를 통한 탐색 결과

그림 3(c)를 보면 “폴리에스터 인조 스웨드의 제조방법” 이라는 발명의 명칭에 대해 앞서 설명한 방식을 적용하여 KIPRIS API를 통해 탐색한 결과 <농색성 및 견뢰도가 우수한 스웨드조 폴리에스터 직물의제조방법>, <나일론해도사 및 일반폴리에스터사를 사용한 스웨드조인공피혁의 제조방법> 와 같이 유사한 특허 명세서가 검색된 것을 확인할 수 있다. 결과적으로 발명의 명칭과 같은 짧은 문장을 활용해서 TF-IDF 방식으로 키워드를 추출할 때, 유사한 선행기술 문헌이 작성되었다.

Ⅲ. 결 론

본 논문에서는 특허 명세서 데이터의 적절한 전처리와 TF-IDF 방식을 활용해 발명의 명칭에서 키워드를 추출하였다. 추출된 키워드를 활용해 KIPRIS Plus-Open API를 통하여 특허 명세서의 선행기술조사 문헌을 적절히 작성하고자 했다. 그 결과 실제 특허 명세서에 작성된 선행기술조사 문헌과 제시한 방식을 활용하여 작성된 선행기술조사 문헌이 유사함을 보였다.

변리사가 본 논문에서 제시한 방식을 선행기술조사 문헌 작성에 활용한다면 직접 수고스럽게 찾아야 하는 기존 방식에 비해 작업의 효율을 높여주고, 시간을 절약시킬 것으로 기대된다.

참 고 문 헌

[1] Inseo Wang, Seokwoo Byun, & Gyun Woo (2021). Building a List of Stopwords for Text Preprocessing. 한국정보과학회 학술발표논문집, 457-459.

[2] Kil, Ho-hyun (2018). The Study of Korean Stopwords list for Text mining. URIMALGEUL : The Korean Language and Literature, 78, 1-25.

[3] Sungjick Lee, & Han-joon Kim (2009). Keyword Extraction from News Corpus using Modified TF-IDF. The Journal of Society for e-Business Studies, 14(4), 59-73.

[4] Young-Hoon Kim, Seung-Min Park, & Dae-Soo Cho (2022). Design of Document Suggestion System based on TF-IDF Algorithm for Efficient Organization of Documentation. Proceedings of the Korean Society of Computer Information Conference, 527-528.