특허명세서의 한국표준산업분류 (KSIC) 코드 자동 분류를 위한 다중 라벨 분류 모델

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

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

A Multi-Label Classification Model for Automatic Classification of Korean Standard Industrial Classification (KSIC) Codes of Patent Specifications

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

요 약

21 세기, 지식기반사회로 접어들면서 특허의 중요성이 커짐에 따라 특허 출원 건수는 꾸준히 증가하고 있으며, 이에 따른 정확한 특허 분류의 중요성이 부각되고 있다. 특허 분류가 부여되는 과정을 인간이 수행하면서 발생하는 비효율성을 개선하고자 본 연구는 특허데이터를 기반으로 한 Multi-Label Classification(다중 라벨 분류) 모델을 활용하여 한국표준산업 분류(KSIC) 코드를 자동으로 지정하는 방법을 제안한다. 17 만 7 천 건의 특허데이터를 수집해 모델을 학습시키고, 발명의 명칭, 요약, 청구범위를 입력으로 각 특허명세서에 해당하는 KSIC 코드를 자동으로 분류하여 제시하였다. 제안된 모델은 KSIC 코드를 지정하는 데 77.98%의 정확도를 보였다.

Ⅰ. 서 론

21 세기에 접어들며 정보와 지식이 새로운 형태의 에너지와 자본으로서 중요도가 높아지는 지식기반사회로 접어들며 가치 창출의 중심이 지식이 되어 지식의 생산, 관리, 활용이 중요한 사회가 되었다[1]. 이에 따라 다음과 같은 현상이 나타났다.

첫째, 급격한 기술의 발전으로 디지털 정보의 양이 급격히 증가하면서 정보 과잉 현상이 발생하였다. 이로 인해 분류 정보의 중요성이 강조되어 통계청이 경제, 산업 등 다양한 분야의 통계 수집과 분석을 목적으로 한국표준산업 분류(KSIC)를 제정하였다[2].

둘째, 새로운 형태의 자본인 지식 재산권 창출에 대한 중요도가 높아지면서 지식 재산권의 대표적인 형태인 특허의 출원 건수가 매년 증가하고 있다.

특허의 분류는 특허에 관한 정보를 이용하는 것을 목적으로 하는 자에게 특허문헌에 대한 접근성을 높이며 심사관의 검색을 쉽게 하도록 기술을 세분화하는 역할을 한다. 특허청의 '특허·실용신안 심사기준'[3]에 따르면 특허 분류는 그림 1 과 같은 절차를 거친 후 부여된다. 현재 특허 분류를 부여하는 절차를 모두 인간이 수행하고 있어 복잡하고 비효율적이다. 이 절차를 간단하고 효율적으로 개선하기 위해서는 이 절차를 대신 수행할 수 있는 모델의 개발이 중요하다[4]. 따라서 본 논문에서는 특허 분류 방식의 실용성과 효율성을 증진시키기 위하여 특허 데이터 기반 Multi-Label Classification(다중 라벨 분류) 모델을 적용해 각 특허명세서에 해당하는 KSIC 코드를 자동으로 지정해 제시하고자 한다.

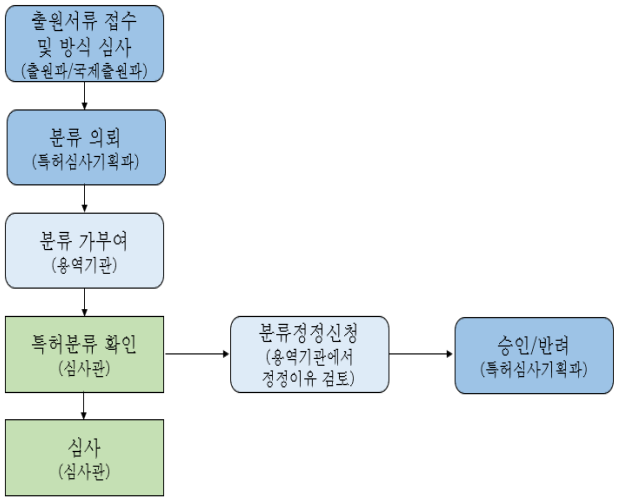

그림 1. 특허 분류 부여 절차

Ⅱ. 본 론

다중 분류 모델은 대표적으로 Multi-Class Classification(다중 클래스 분류)와 Multi-Label Classification 이 존재한다. Multi-Class Classification 의 경우 각 샘플이 단 하나의 label 을 가지며 각 label 은 유일한 class 를 나타내는 방식으로 다중의 정답이 존재할 수 없다. 하지만 특허 분류에서는 한 특허 문서가 하나의 특허 분류 코드만을 가지지 않고 다중 특허 분류 코드를 가지는 경우에 넓은 기술 영역에 대한 보호를 받을 수 있고 시장에서의 다양한 활용 가능성을 시사해 출원자에게 더 큰 시장 영향력을 제공할 수 있으며 특허 검색을 용이하게 한다는 이점이 있다. Multi-Class Classification 과 달리 Multi-Label Classification 은 각 샘플이 여러 개의 label 을 가질 수 있어 다중의 정답이 동시에 존재하는 방식이므로 여러 개의 특허 분류 코드 제시가 가능하다. 따라서 본 연구에서는 분류 모델로서 Multi-Label Classification 을 선정하였다.

그림 2 는 본 연구의 전반적인 수행 과정을 나타낸 것이다. 특허명세서를 KSIC 코드로 분류하기 위해 KSIC 코드 데이터를 수집하여 Multi-Label Classification 모델을 파인튜닝(Fine-tuning) 하였다. 발명의 명칭, 요약, 청구범위를 입력하여 KSIC 코드를 분류한 후 검증과정을 통해 정확도를 확인하였다.

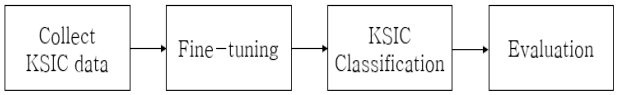

그림 2. 연구 수행 과정

특허명세서를 KSIC 코드로 분류하기 위해 특허데이터 17 만 7 천 건을 KEYWERT 에서 웹 크롤러를 이용하여 추출하였다. 추출한 특허데이터는 특허명세서의 발명의 명칭, 요약, 청구범위 항목들과 각 특허명세서에 대응하는 KSIC 코드 A, B, C, D, E, J 로 이루어져 있다. KSIC 에서 대문자 알파벳은 산업분야의 대분류를 나타내며, 추출한 데이터가 각각 나타내는 산업분야의 대분류와 대분류명을 표 1 에 정리하였다.

표 1 KSIC 코드의 대분류와 대분류명

대분류	대분류명
A	농업, 임업, 어업
B	광업
C	제조업
D	전기, 가스, 증기 및 공기 조절 공급업
E	수도, 하수 및 폐기물 처리, 원료 재생업
J	정보통신업

수집한 데이터를 분류 모델을 학습시키는 데 사용하기 위해서는 정제 작업을 수행해야 한다. KoELECTRA 언어 모델을 이용하여 모델이 이해할 수 있도록 문장을 단어 단위로 분해하는 토큰화를 하였다. 발명의 명칭, 요약, 청구범위를 입력하였을 때 KSIC 코드를 알려주는 것을 목적으로 분류모델을 학습시키기 위한 과정을 수행하고 있으므로 수집한 데이터를 바탕으로 발명의 명칭, 요약, 청구범위를 입력, KSIC 코드를 출력으로 데이터를 분류하였다. 텍스트 형태로 구성된 코드를 고정된 길이의 수치형 벡터로 변환하기 위해 임베딩 과정을 수행하였다. 임베딩 과정에서는 각 토큰에 해당하는 고유한 하나의 정수 값, 즉 인덱스로 토큰을 매핑하여 변환한다.

정제 작업을 거친 데이터의 70%는 훈련 데이터로, 나머지 30%는 실험과 검증을 위한 데이터로 분리하였다. 훈련 데이터를 이용해 총 20 epoch 의 학습을 수행하여 Multi-Label Classification 모델을 특허명세서를 기반으로 KSIC 코드를 지정하는 데 특화시켰다.

실험 및 검증에 사용한 데이터는 학습에 이용한 데이터와 동일한 정제 작업을 거쳤으며 사용자에게 다양성을 제공하고자 3 개의 결과를 제시한다. 결과의 평가지표로 F1-score 방식을 이용하였으며, 해당 방식으로 계산한 결과 별 정확도는 표 2 에 정리하였다. 결과는 가장 유사도가 높은 순서대로 제시되므로 결과 1 이 약 47.32%로 가장 높은 정확도를 보인다. 제시한 3 가지 결과에 대한 총 정확도는 77.98%로 우수한 성능을 보였다.

표 2 대분류 A, B, C, D, E, J 에 대한 분류 코드 정확도

정확도 [%]	결과 1	47.32
	결과 2	19.63
	결과 3	11.03
	총 정확도	77.98

Ⅲ. 결 론

본 논문에서는 특허 분류 부여 절차를 효율적으로 개선하기 위해 특허명세서의 KSIC 코드를 Multi-Label Classification 모델을 이용하여 자동으로 분류해 제시하고자 하였다. 이를 위해 17 만 7 천 건의 특허명세서 기반 데이터를 추출하였으며 모델 학습에 이용하기 위해 추출한 데이터를 정제하는 과정을 수행하였다. 정제한 데이터의 70%는 모델 학습에 이용하였고 30%는 실험 및 검증을 하는 데 이용하였다. 발명의 명칭, 요약, 청구범위를 입력으로 하여 KSIC 코드로 분류된 결과 3 가지를 제시하도록 진행하였으며, F1-score 방식으로 측정한 결과 모델의 정확도는 77.98%로 우수한 성능을 보임을 확인하였다.

참 고 문 헌

[1] 이재성, 전승표, 유형선. (2018). 한국표준산업분류를 기준으로 한 문서의 자동 분류 모델에 관한 연구. 지능정보연구, 24(3), 221-241.

[2] 김명선, 한동희. (2022.12). 기술문서의 한국표준산업분류 자동분류를 위한 특허기반 BERT 모델. 한국정보과학회 학술발표논문집.

[3] "Guidelines for Examination," Korean Intellectual Property Office, ISSN 2092-8866.

[4] 임소라, 권용진. (2017). 특허문서 필드의 기능적 특성을 활용한 IPC 다중 레이블 분류. Journal of Korean Society for Internet Information, 18(1), 77-88.