특허문서의 IPC 분류를 위한 데이터 변환 및 통합

박수현[1], 김진[2]
[1]상명대학교 지능·데이터융합학부 학부생
[2]상명대학교 빅데이터융합전공 교수
202110794@sangmyung.kr, jinkim@smu.ac.kr

Pre-processing for IPC Classification of Patent Documents

Su-Hyun Park[1], Jin Kim[2]
[1]Faculty of Artificial Intelligence and Data Engineering, Sangmyung University
[2]Big Data Convergence Major, Sangmyung University

요 약

4차 산업혁명으로 다양한 기술과 아이디어가 생겨나고 있고, 이를 보호하기 위한 특허는 그 등록 건수가 매년 증가하는 추세이다. 그러나 현재 특허문서를 분류하는 과정을 수동으로 진행하고 있기에 이를 자동으로 진행할 수 있는 분류기를 생성할 필요를 느꼈고, 본 논문에서는 특허문서를 분류기에 적용할 데이터의 전처리 과정 중 데이터 변환과 통합 과정을 다루었다.

1. 서론

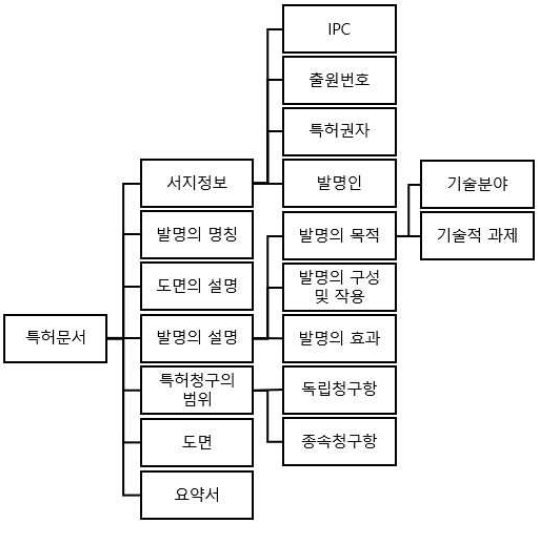
4차 산업혁명을 통해 수많은 형태의 데이터를 활용할 수 있게 되면서, 기존보다 향상된 기술과 아이디어들이 발생하고 있다. 이와 동시에 자신의 아이디어나 기술 등을 보호할 목적으로 특허 등록 건수 또한 급등하는 추세이다. 실제로 우리나라의 2012년부터 2022년 동안 지재권 출원 특허 수는 총 160,118,300건에 달하며, 동기간 내 특허는 연평균 4.2%의 성장률[1]을 보였다. 특허 등록 비율이 높아지는 것과 기술의 발전 추세를 고려한다면 특허 등록 비율은 지속적으로 늘어날 것이고, 동시에 특허문서를 체계적이고 빠르게 분류하는 기술이 요구되고 있다. 그러나 현재 특허문서 분류 과정에는 사람이 직접 참여하여 분류를 진행하고 있어 시간이 많이 소요되고 있다. 이러한 과정을 자동으로 바꾼다면 효율적으로 특허문서 분류를 진행할 수 있을 것이다.

특허문서의 분류 과정을 자동으로 진행하기 위해서는 성능이 좋은 분류기가 필요하다. 또한, 분류기 학습에 사용할 충분한 데이터도 필요하므로, 본 논문에서는 특허문서의 IPC 분류를 위한 효율적인 데이터 전처리 과정을 다루었다.

2. 특허문서와 IPC 구조

특허로 공개된 기술을 활용하기 위해서는 기술을 찾고 파악하기 위한 IPC가 필요하고, 결국 IPC의 정확하고 신속한 분류가 필수로 필요하다.

특허문서의 구조는 다음과 같이 서지정보, 발명의 명칭, 도면의 설명, 발명의 설명, 특허청구의 범위, 도면, 요약서로 구성된다. 특허문서를 분류하는 내용의 IPC는 서지정보 안에 포함되어 있다.



(그림 1) 특허문서의 구조.

특허문서를 분류하는 코드인 IPC는 International Patent Classification의 약자로, 국제특허분류를 의미한다. 이는 특허문헌에 포함된 기술 및 권리정보에 용이하게 접근할 수 있도록 하는 코드이다.

IPC의 구조는 다음과 같이 총 5단계의 계층 구조로, 상위부터 8개의 섹션(Section), 128개의 클래스(Class), 약 650개의 서브클래스(Subclass), 약 6,800개의 메인그룹(Maingroup), 약 65,000개 이상의 서브그룹(Subgroup)[2] 순으로 이루어져 있다.

섹션(Section)	A	B	C	D	E	F	G	H
클래스(Class)	F01	F02	....	F45	F46	...	F42	F99
서브클래스(Subclass)	F01B	F01C	...	...	F01L	...	...	F01P
메인그룹(Maingroup)	F01C1/00	F01C3/00	....	....	....	...	F01C19/00	F01C20/00
서브그룹(Subgroup)	F01C1/02	F01C1/04	....	F01C1/44	F01C1/46	...	F01C1/352	F01C1/356

(그림 2) IPC의 구조.

3. 사용한 데이터와 전처리 과정

이번 전처리에 사용한 데이터는 크게 두 가지로, 심사진행상태, 출원번호, 발명의명칭, IPC분류 등이 포함된 엑셀 파일과, 특허문서의 데이터가 포함된 xml 데이터이다.

우선 '특허문서 필드의 기능적 특성을 활용한 IPC 다중 레이블 분류'[2]를 참고하여, 제목, 요약, 청구항, 기술분야, 배경기술, 독립청구항이 포함된 데이터를 얻는 것을 목적으로 설정했다.

전처리 과정은 다음과 같이 크게 세 과정으로 나누어진다. 엑셀 데이터 전처리가 첫 번째, xml 데이터 전처리가 두 번째, 전처리 한 엑셀 데이터와 xml 데이터를 합치는 것이 마지막이다.

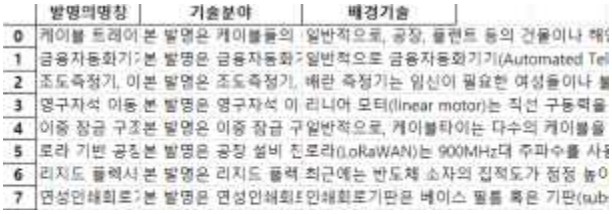
엑셀 데이터는 R 환경에서 '발명의명칭', 'IPC분류', '요약', '청구항'만 추출 후 전처리를 진행했다.. '발명의명칭'은 한글로 된 명칭만 남겨두었고, 'IPC분류'는 날짜를 삭제하고 구분자를 ','(쉼표)로 바꾸었다. '요약'에서는 괄호와 괄호 안의 내용을 삭제하고, '청구항'에서는 독립청구항 내용만 남겨두었다. xml 데이터는 python 환경에서 '발명의명칭', '기술분야', '배경기술' 추출 후 전처리를 진행했다. '발명의명칭'은 한글로 된 명칭만 남겨두었고, '기술분야'와 '배경기술'은 추가 설명이나 특수기호는 전부 삭제한 후, 엑셀 파일로 저장했다. 마지막으로, 앞서 전처리 한 엑셀 파일과 xml 파일을 python 환경에서 '발명의명칭'을 기준으로 합쳐, 새로운 엑셀 파일을 만드는 과정으로 진행했다.

4. 전처리 결과

엑셀 데이터의 전처리 결과는 다음과 같다. 전처리 되어 총 12개의 파일로 엑셀로 추출했다.

(그림 3) 전처리 한 엑셀 파일의 일부.

이어, xml 데이터의 전처리 결과는 다음과 같다. 전처리 후 총 14개의 엑셀 파일로 추출하였다.



	발명의명칭	기술분야	배경기술
0	케이블 트레이	본 발명은 케이블들의	일반적으로, 공장, 플랜트 등의 건물이나 해
1	금융자동화기기	본 발명은 금융자동화	일반적으로 금융자동화기기(Automated Tel
2	조도측정기, 이	본 발명은 조도측정기,	배란 측정기는 임신이 필요한 여성들이나 불
3	영구자석 이동	본 발명은 영구자석 이	리니어 모터(linear motor)는 직선 구동력을
4	이중 잠금 구조	본 발명은 이중 잠금 구	일반적으로, 케이블타이는 다수의 케이블을
5	로라 기반 공동	본 발명은 공장 설비	진로라(LoRaWAN)는 900MHz대 주파수를 사용
6	리지드 플렉서	본 발명은 리지드 플렉	최근에는 반도체 소자의 집적도가 점점 높아
7	면성인쇄회로기	본 발명은 면성인쇄회로	인쇄회로기판은 베이스 필름 혹은 기판(sub

(그림 4) 전처리 한 xml 파일의 일부.

마지막으로, 앞서 전처리 한 엑셀 데이터와 xml 데이터를 합친 결과이다. 총 12개의 파일의 엑셀 데이터로 추출하였다.

	발명의명칭	IPC분류	요약	독립청구항	기술분야	배경기술
0	선박	B63J3/04, E	선박을 제공한	[독립청구항] 선	본 발명은 선박	비교적 규모가 큰
1	선박	B63J3/04, E	선박을 제공한	[독립청구항] 선	본 발명은 선박	비교적 규모가 큰
2	선박	B63J3/04, E	선박을 제공한	[독립청구항] 선	본 발명은 선박	비교적 규모가 큰
3	선박	B63B17/00	선박을 제공한	[독립청구항] 선	본 발명은 선박	비교적 규모가 큰
4	선박	B63J3/04, E	선박을 제공한	[독립청구항] 선	본 발명은 선박	비교적 규모가 큰
5	선박	B63J3/04, E	선박을 제공한	[독립청구항] 선	본 발명은 선박	비교적 규모가 큰
6	선박	B63J3/04, E	선박을 제공한	[독립청구항] 선	본 발명은 선박	비교적 규모가 큰
7	선박	B63J3/04, E	선박을 제공한	[독립청구항] 선	본 발명은 선박	비교적 규모가 큰
8	선박	B63J3/04, E	선박을 제공한	[독립청구항] 선	본 발명은 선박	비교적 규모가 큰
9	선박	B63B79/30	선박을 제공한	[독립청구항] 선	본 발명은 선박	비교적 규모가 큰
10	선박	G08B21/02	선박을 제공한	[독립청구항] 선	본 발명은 선박	비교적 규모가 큰

(그림 5) 엑셀과 xml 파일을 합친 파일의 일부.

5. 향후 연구 방향

본 연구를 바탕으로 특허문서 분류 과정을 자동화할 수 있는 분류기를 생성하는 것이 목적이다.

따라서, 해당 전처리를 활용하여 특허문서의 IPC를 분류하는 여러 방식의 분류기 생성이 첫 번째, 생성한 여러 분류기의 성능을 비교하여 가장 우수한 분류기를 판별하는 것이 두 번째 과제이다.

참고문헌

[1] 특허청, 한국특허정보원, 2022 통계로 보는 특허동향, 2022,

[2] 임소라, 권용진, 특허문서 필드의 기능적 특성을 활용한 IPC 다중 레이블 분류, 한국 인터넷 정보학회, 18권, 1호, 79~82쪽