특허문서의 IPC 분류기 생성을 위한 데이터 전처리

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

Data Pre-processing for Create IPC Classifiers for Patent Documents

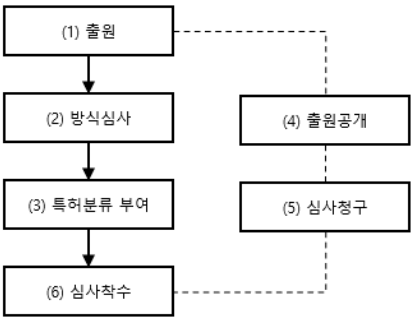
Su-Hyun Park[1], Jin Kim[2]
[1]Big Data Convergence Major, Sangmyung University
[2]Big Data Convergence Major, Sangmyung University

요 약

특허심사절차는 짧지 않은 과정으로 이루어져 있는데, 현재 모든 절차가 사람이 직접 관여하여 진행되고 있다. 특허심사절차의 효율적 시간 분배를 위해, 특허문서 분류 과정의 자동화 처리 필요성을 느끼게 되었다. 따라서, 본 논문에서는 해당 분류기 생성을 위한 데이터의 전처리 과정을 다루었다.

1. 서론

특허청의 특허심사절차 일반[1]에 따르면, 특허심사 과정 중 특허문서의 분류 과정은 아래의 절차에 따라 진행된다.

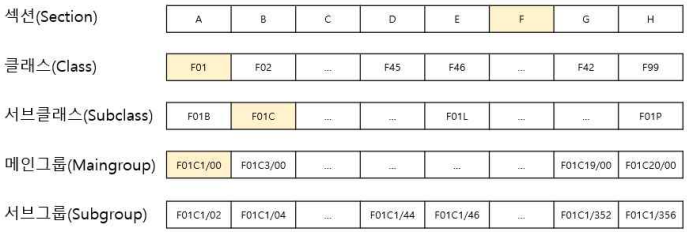


(그림 1) 특허문서의 분류 과정.

특허문서의 분류 과정을 지나면 특허결정, 의견서/보정서, 재심사 청구 등의 과정을 거쳐 심판/소송/판결 단계까지 진행된다. 이처럼 특허심사절차의 모든 과정에 사람이 직접 관여하여, 오랜 시간이 소요된다.

이러한 배경에서 효율적인 절차 진행에 대해 고민하게 되었고, 특허문서의 분류 과정을 자동화 하는 분류기의 필요성을 느끼게 되었다.

다만, 특허문서를 분류하는 코드인 International Patent Classification (IPC)의 구조는 (그림 2)[2]의 구성처럼 방대한 분류로 이루어져 있기 때문에, 서브그룹까지 판단하는 분류기를 생성하기에는 현실적인 어려움이 있을 것으로 예상된다.



(그림 2) IPC의 구조.

따라서 IPC의 구조 중 서브클래스까지 판단하는 분류기 생성을 목표로, 본 논문에서는 해당 분류기 생성을 위한 데이터 전처리에 대해 논할 예정이다.

2. 사용 데이터와 데이터 전처리

특허문서 분류의 자동화를 위한 분류기에 사용할 데이터는, 기존의 특허문서 데이터가 포함된 xml 형식의 데이터이다.

'특허문서 필드의 기능적 특성을 활용한 IPC 다중 레이블 분류'[3]를 보면, 청구항 필드는 타 필드에 비해 데이터가 많아 분류기 성능을 떨어트리는 요소로 작용하는 것을 알 수 있다. 다만, 청구항 중 독립청구항만 포함했을 경우의 분류기 성능은 확인되지 않았다.

따라서 독립청구항 여부에 따른 분류기 성능 비

교를 위해, 본 논문에서는 발명의 명칭, 요약, 기술분야, 배경기술, IPC가 포함된 첫 번째 데이터와, 첫 번째 데이터에서 독립청구항을 추가한 두 번째 데이터를 얻는 과정을 진행했다.

3. 전처리 과정과 결과

전처리는 크게 세 가지 과정으로 진행했다.

첫 번째 과정에서는 파이썬 환경에서 xml 파일을 불러온 후, 발명의 명칭, IPC, 요약, 독립청구항, 기술분야, 배경기술을 추출하여 엑셀 형식으로 저장하였다. 추출하는 과정에서 필요 없는 기호 등을 제거하였으며, 14개의 파일을 추출하였고 그 결과는 (그림 3)과 같다.

	발명의명칭	IPC	요약	독립청구항	기술분야	배경기술
0	케이블 트레	H02G 3/04	본 발명은	제1 케이블 트	본 발명은	일반적으로, 공장, 플랜트 등의 건물이나 해
1	금융자동화기	G07F 19/0	본 발명은	금융거래를 수	본 발명은	일반적으로 금융자동화기기(Automated Tell
2	조도측정기,	A61B 10/0	본 발명은	타액의 황체	본 발명은	배란 측정기는 임신이 필요한 여성들이나 불
3	영구자석 이	H02K 11/2	본 발명은	복수 개의 주	본 발명은	리니어 모터(linear motor)는 직선 구동력을
4	이중 잠금 구	B65D 63/1	본 발명의	기다란 스트랩	본 발명은	일반적으로, 케이블타이는 다수의 케이블을
5	로라 기반 공	G08C 17/0	본 발명은	공장의 각 기	본 발명은	로라(LoRaWAN)는 900MHz대 주파수를 사용
6	리지드 플렉	H05K 3/46	본 발명은	a)제1내층회로	본 발명은	최근에는 반도체 소자의 집적도가 점점 높아
7	연성인쇄회로	H05K 3/00	시트 내에	하나의 시트	본 발명은	인쇄회로기판은 베이스 필름 혹은 기판(subs

(그림 3) 필요한 필드를 추출한 xml 파일의 일부.

두 번째 과정에서는 첫 번째 과정에서 추출한 14개의 엑셀파일을 파이썬 환경에서 불러와 값이 들어있지 않은 행을 삭제하였다. 동시에 14개의 엑셀 파일을 하나의 새로운 엑셀 파일로 추출하였다. 해당 과정은 추후 학습할 모델에 성능 저하를 발생시키는 오류를 제외하는 목적이 담겨있으며, 최종적으로 추출된 엑셀 파일은 (그림 4)와 같다.

	발명의명칭	IPC	요약	독립청구항	기술분야	배경기술
0	케이블 트레	H02G 3/04	본 발명은	제1 케이블	본 발명은	일반적으로, 공장, 플랜트 등의 건물이나 해
1	금융자동화기	G07F 19/0	본 발명은	금융거래를	본 발명은	일반적으로 금융자동화기기(Automated Te
2	조도측정기,	A61B 10/0	본 발명은	타액의 황체	본 발명은	배란 측정기는 임신이 필요한 여성들이나 불
3	영구자석 이	H02K 11/2	본 발명은	복수 개의 주	본 발명은	리니어 모터(linear motor)는 직선 구동력을
4	이중 잠금 구	B65D 63/1	본 발명의	기다란 스트랩	본 발명은	일반적으로, 케이블타이는 다수의 케이블을
5	로라 기반 공	G08C 17/0	본 발명은	공장의 각 기	본 발명은	로라(LoRaWAN)는 900MHz대 주파수를 사용
6	리지드 플렉	H05K 3/46	본 발명은	a)제1내층회로	본 발명은	최근에는 반도체 소자의 집적도가 점점 높아
7	연성인쇄회로	H05K 3/00	시트 내에	하나의 시트	본 발명은	인쇄회로기판은 베이스 필름 혹은 기판(sub

(그림 4) NONE값을 제거한 엑셀 파일의 일부.

세 번째 과정에서는, 두 번째 과정에서 추출한 엑셀파일로부터 명사를 추출하였다. 해당 과정은 분류기 학습에 데이터를 사용하기 위해, 일관적인 형태이자 의미가 있는 '명사'만을 추출할 필요성을 느껴 진행하게 되었다. 파이썬 환경에서 한국어 데이터 전처리에 사용되는 KoNLPy 패키지의 Kkma 라이브러리를 활용하여 명사를 추출하였다. 동시에 독립청구항이 분류기 성능에 미치는 영향을 판단하기 위해, 독립청구항이 포함된 파일과 독립청구항이 포함되지 않은 파일로 분할하였다. 그 결과는 (그림 5), (그림 6)과 같다.

	발명의명칭	요약	독립청구항	기술분야	배경기술
0	케이블, 트레	발명, 케이	제, 제1, 1, 케	발명, 케이	일반적, 공
1	금융, 자동화	발명, 금융	금융, 금융기	발명, 금융	일반적, 금
2	조도, 측정기	발명, 조도	타액, 황체,	발명, 조도	배란, 측정
3	영구, 영구자	발명, 영구	복수, 개, 주	발명, 영구	리니어, 모
4	이중, 이중 잠	발명, 일실	스트랩타이	발명, 이중	일반적, 케
5	로라, 기반, 공	발명, 공장	공장, 기계,	발명, 공장	로라, 주파
6	리지드, 회로	발명, 제, 1	제, 제1, 1, 내	발명, 리지	최근, 반도
7	연성, 인쇄, 회	시트, 내, 피	하나, 시트,	발명, 연성	인쇄, 인쇄
8	드로소필라,	발명, 메르	메르스, 바C	발명, 메르	중동, 호흡

(그림 5) 독립청구항이 포함된 명사 추출 파일.

	발명의명칭	요약	기술분야	배경기술
0	케이블, 트레	발명, 케이	발명, 케이	일반적, 공
1	금융, 자동화	발명, 금융	발명, 금융	일반적, 금
2	조도, 측정기	발명, 조도	발명, 조도	배란, 측정
3	영구, 영구자	발명, 영구	발명, 영구	리니어, 모
4	이중, 이중 잠	발명, 일실	발명, 이중	일반적, 케
5	로라, 기반, 공	발명, 공장	발명, 공장	로라, 주파
6	리지드, 회로	발명, 제, 1	발명, 리지	최근, 반도
7	연성, 인쇄, 회	시트, 내, 피	발명, 연성	인쇄, 인쇄
8	드로소필라,	발명, 메르	발명, 메르	중동, 호흡

(그림 6) 독립청구항이 포함되지 않은 명사 추출 파일

4. 향후 연구 방향

본 논문에서는 명사 추출까지 진행했으나, 추출한 명사에 불용어가 포함되어 있을 가능성이 있다. 때문에, 제외할 불용어의 기준을 정하고 추출한 명사에서 불용어를 제거하는 과정이 필요할 것이다.

최종 목표는 본 연구를 바탕으로 준비된 데이터를 활용하여, 특허문서의 IPC 분류 자동화를 위한 분류기를 생성하는 것이다. 최종 목표를 달성하기 위해 사용 가능한 모델들 중, 필자가 고려한 모델은 Support Vector Machine (SVM), K Nearest Neighbor (KNN), RandomForest이다.

또한 위에서 언급했듯이, 기본적으로 성능 저하에 영향을 미치는 청구항 중 독립청구항만 포함하였을 때의 성능 평가가 필요하다. 따라서, 본 연구에서 분할한 두 가지 데이터를 가지고 분류기를 생성하여 성능을 평가하는 것도 필요하다고 보여진다.

참고문헌

[1] 특허청, 한국특허정보원, 심사절차일반. 1쪽

[2] 박수현, 김진, "특허문서의 IPC 분류를 위한 전처리", ACK 2023 학술발표대회 논문집, 30권, 2호, 401쪽

[3] 임소라, 권용진, "특허문서 필드의 기능적 특성을 활용한 IPC 다중 레이블 분류", 한국 인터넷 정보학회, 2017, 18권, 1호, 86쪽