기술문서 분류를 위한 통계기반 기계학습 모델 성능비교 및 한계 연구

김진구, 유헌창
고려대학교 컴퓨터정보통신대학원 빅데이터융합학과
kim_jin_gu@korea.ac.kr, yuhc@korea.ac.kr

Performance Comparison of Statistics-Based Machine Learning Model for Classification of Technical Documents

Jin-gu Kim, Heonchang Yu
Dept. of Computer & Information Technology, Korea University

요 약

본 연구는 국방과학기술 분야의 특허 및 논문 실적을 이용하여 통계기반 기계학습 모델 4 종을 학습하고, 실제 분석 대상기관의 데이터 입력결과를 분석하여 실용성에 대한 한계점 분석을 목적으로 한다. 기존 연구에서는 특허분류코드를 기준으로 분류하여 특수 목적으로 활용하거나 세부 연구 범위 내 연구 주제탐색 및 특징연구 등 미시적인 관점에서의 상세연구 활용 목적인 반면, 본 연구는 거시적인 관점에서 연구의 전체적인 흐름과 경향성 파악을 목적으로 한다. 이에 ICT 기술 138 종의 특허 및 논문 30,965 건과 국방과학기술 192 종의 특허 및 논문 23,406 건을 학습데이터로 각 모델을 학습하였다. 비교한 통계기반 학습모델은 Support Vector Machines, Decision Tree, Naïve Bayes, XGBoost 모델이다. 학습데이터에 대한 학습검증 단계에서는 최대 99.4%의 성능을 보였다. 다만, 실제 분석대상기관의 특허 및 논문 12,824 건으로 입력분석한 결과, 모델별 편향성 문제, 데이터 전처리 이슈, 다중클래스 및 다중레이블 문제를 확인, 도출한 문제에 대한 해결방안을 제시하고 추가 연구의 방향성을 제시한다.

1. 연구배경 및 목적

기술기반의 기관 및 기업들은 변화하는 시장 환경에 대응함과 동시에 미래를 선도하기 위해서 시장에서의 기술융합 흐름 파악과 기술 전략 수립을 위한 정보분석에 많은 노력을 기울이고 있다. 타분야의 연구형태와 방식이 또다른 연구에 긍정적인 영향을 주어 융합형태의 발전을 가속화되며[1], 융합에 대한 분석방식도 연구가 활발히 이루어지고 있다[2]. 특히 디지털기술로 촉발된 초연결 기반의 지능화 핵심기술인 인공지능, 빅데이터, 사물인터넷, 지능형 로봇 등 4 차 산업혁명 기술(이하 4IR 기술)은 산업 전반에 기술융합 및 개발방식의 혁신에 영향을 주고 있어서, 시장환경에서 추적관리해야 할 기술분야 중 대표적인 그룹으로 식별되고 있다.

본 연구의 거시적인 목적은 기술융합의 분석에 있어서 4IR 기술이 타 국가연구개발기술에 미치는 영향력 및 연구방식의 변화를 분석하는데 있다. 일종의 매트릭스 분석법으로서, 4IR 기술그룹을 세로축으로 정렬하고, 분석대상 그룹(예를 들어, 국방과학기술 그룹)을 가로축으로 정렬한다. 그리고 각 세부분야의 기술문서(특허, 논문, 보고서 등)의 연구주제에 대한 누적분석을 상호 매칭 방식으로 분석하고, 전체의 시기별 분석으로 통해서 기술융합의 형태와 흐름, 그리고 아직 시도되지 않은 기술융합의 기회를 발견하는데 있다.

전체 연구목적을 달성하기 위해서 각 기술 그룹의 관련 연구에 대한 명확한 분류가 필요하다. 본 연구를 전체연구의 진행 전, 4IR 기술 그룹 및 타 국가과학기술분야의 기술문서를 대상으로 각 그룹별 기계학습 모델 구현을 통한 문서 분류의 성능을 비교하여 최적의 기술문서 분류를 수행하는데 있다.

2. 관련 연구

기존 분류모델 연구에서는 특허분류를 위해서 다량의 특허 클래스를 모으고, 비교적 적은 수의 학습데이터를 이용하는 대신 다양한 딥러닝 모델 및 특허 필드의 항목을 조합하여 분류기의 성능을 높이고자 하였다[3]. 이 연구에서는 분류기의 학습데이터에 제목, 요약, 전체청구항 및 특허 분류 코드인 IPC 를 포함하고 있어 모델의 학습에 직접적인 영향을 주었다.

정답이 포함된 데이터셋으로 학습이 이루어진 것과 같으므로 실용성은 낮다. 다른 분류연구에서는 특허 문헌의 필드인 발명의 명칭, 요약, 배경기술, 기술분야 및 독립청구항의 조합을 통한 문서분류 성능을 비교하였다. [4]. 이를 통해 학습된 분류모델에서도 기술분야가 학습에 포함되고 있어서, 실제 활용에서 정보가 주어지지 않는 문서에 대한연관 특허 분류로는 활용성이 낮게 된다. 다른 형태로는 분류 대상 범위를 세부 주제분야로 한정하고 지도학습으로 모델을 학습하고 개발하는 방식으로서[5], 기술분류체계가 넓은 분야를 대상으로 사용하는 것에는 한계가 있다. 본 연구에서는 보편적인 문서 분류로서의 활용성을 목적으로 ICT 기술분야 138 종의 분류와 국방과학기술분야 192 종의 분류를 위해 목표로 한다.

분류모델의 비교대상은 대상 데이터 집합을 바탕으로 새로운 데이터가 속할 카테고리를 판단하는 비확률적 이진 선형 분류 모델인 Support Vector Machines, 의사결정 규칙과 그 결과에 대한 트리구조를 찾는 Decision Tree, 특성들 사이의 독립을 가정한 Bayes 정리를 적용한 확률분류기인 Naïve Bayes, 그리고, 여러 개의 약한 Decision Tree 를 조합해서 사용하는 앙상블 기법의 하나인 XGBoost(eXtreme Gradient Boost)를 학습하여 성능을 비교하였다.

3. 연구 절차

본 연구의 진행을 위해 학습데이터 수집, 기술적 데이터 분석 및 전처리, 탐색적 데이터 분석, 예측적 데이터 분석, 모델 결과 평가 및 피드백의 단계를 수립하였다.

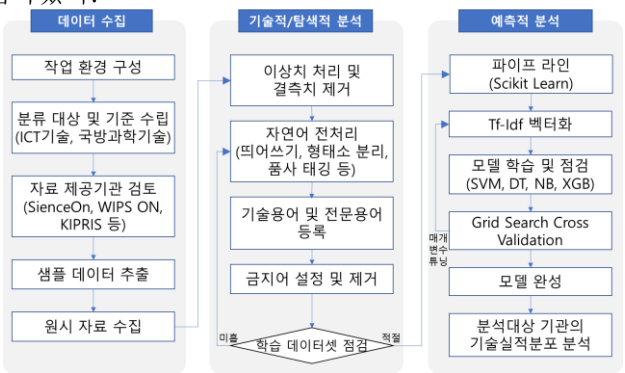

그림 1. 연구 절차 순서도

3.1 학습데이터 수집

문서분류 모델의 학습을 위한 데이터 수집의 단계로서, 한국과학기술정보연구원에서 운영하는 과학기술 지식인프라 통합서비스인 사이언스온을 통하여 특허 및 논문 자료를 수집하였다. 자료의 수준은 명칭, 저자, 발행년도, 초록(요약), 발행기관 등의 기본적인 메타데이터 정보에 해당하며, 원문 전체는 수집하지 않았다. 이유는 논문의 경우 각 학회 및 저자 개인과 별건 협약을 통해서만 원문 전체 제공이 가능하며, 다양한 사유로 저작권 이슈가 있다.

원시데이터에 대한 레이블링 작업은 기술분류의 기준이 학습 레이블링과 동일하게 반영되므로 객관적인 기준에 따라 자료를 수집하고자, 정부 부처의 기술관련 발간서의 기준을 따랐다[6]. 또한, 전체연구의 목적이 ICT 기술의 타 기술에 대한 기술융합 연구임을 감안하여 영향을 발휘한 기술 축에는 인공지능, 사물인터넷, 빅데이터, 클라우드 등의 항목을 4 차 산업혁명 선도기술(이하 4IR 기술) 138 종을 기준하였으며, [7] 피영향 기술 축에는 센서, 정보통신, 제어전자, 탄약/에너지, 추진, 화생방, 소재의 체계적 분류된 국방과학기술분류체계의 기술항목(이하 국방과학기술) 192 종을 기준으로 키워드 검색을 통해 수집하였다[8].

3.2 기술적 데이터분석 및 전처리

수집한 원시데이터는 4IR 기술에서 특허 21,780 건, 논문 13,313 건이며, 국방과학기술은 특허 9,768 건, 논문 21,352 건이다. 학습은 4IR 기술그룹과 국방과학기술 그룹 두가지를 별건으로 구분하여 학습하였다. 학습에는 특허의 발명명칭, 요약항을 사용하며, 논문은 논문명과 초록을 대상으로 학습하게 된다. 메타데이터 중 특허의 발명자, 발행기관 및 논문의 발행처, 논문저자, 출처(학회명) 등은 모델의 학습 가중치를 훼손할 수 있으므로 제외하였다. 선별한 데이터를 사용하여 한국어 자연어처리의 전처리 과정에 따라 기술적 데이터분석 및 전처리를 수행하였다. 결측치 제거, 띄어쓰기 보정, 음절분리, 형태소 분리, 품사 태깅, 기술용어 및 전문용어 통일화, 어휘사전에 기술용어 및 전문용어 등록, 금지어 제거 과정을 수행하였다.

3.3 데이터셋 검증

정제된 데이터가 기술분류에 따른 레이블링 적절성을 살피고자 각 분류된 전처리 데이터의 워드크라우드를 도출하고, 전문가 분석을 통해 각 기술분류의 적절성을 판단하였다. 레이블링 과정에서 다중 레이블이 적용되는 기술에 대해서는 유사성이 높다고 판단하는 하나의 기술만 적용하였다. 또한, 식별된 용어가 특허 및 논문에서 일반적으로 쓰이는 용어, 미처 걸러내지 못한 금지어 등에 해당하는 경우, 금지어 등록 및 기술용어 및 전문용어 등록을 재진행하여 3.2 단계와 3.3 단계를 반복적으로 수행하여 학습데이터를 정제하였다. 또한, Tf-Idf 처리를 여러 문서에 반복적으로 등장하는 용어의 중요도를 낮추고, 각 기술분류에서 비중있게 등장하는 단어의 가중치를 높이도록 하였다.

3.4 예측적 데이터 분석

검증이 완료된 데이터셋은 학습용과 검증용으로 80:20 으로 분할하였으며, 각 학습용 데이터셋에는 특허 및 논문의 제목과 초록에서 추출한 명사형 단어가 포함된다. 학습 후 검증결과 Support Vector Machines(선형타입) 95.8%, Support Vector Machines(다항타입) 75.6%, Decision Tree 97.6%, Naïve Bayes 60.1%, XGBoost 99.4%의 성능을 보였다. SVM 모델에 대한 교차검증 결과 94.9%로 도출되었다. 정밀도, 재현율, F1-score 에 대해서는 Micro 에서는 모두 99.4%, Macro 에서는 각각 88.6%, 88.6%, 88.5%의 결과를 도출하였다.

3.5 실 데이터 입력 분석 평가

완성한 기술분류 모델을 이용하여 분석 대상기관의 특허와 논문의 기술분류 분석을 수행하였다. 대표적인 국방과학기술 연구기관인 국방과학연구소(ADD)의 특허 6,121 건, 논문 6,412 건을 사용하였다. 해당 데이터는 가공되지 않은 원형의 레이블이 없는 데이터이며, 모델의 처리 결과를 통해 기관의 기술연구 흐름을 분석하고자 기대하였다.

그러나 분류 모델의 처리 결과, 학습 시험데이터의 이상적인 결과와는 다른 편향성 문제가 확인되었다. 모델별로 특정 기술 카테고리에 과적합된 결과가 도출되었다. 그림 2 는 국방과학기술 중분류(Lv2)의 결과로서, 모델별로 특정 기술항목에 평향되는 결과를 도출하였다. 이를 통해서 추정할 수 있는 것은 전처리 데이터의 문제보다는 모델별 특성에 따라 결정되는 단어 요소가 있을 것이라는 추정이다. 이유는 전처리된 데이터에서 전체 학습에 영향을 미치는 용어가 있는 경우, 모든 모델에서 동일한 기술항목으로 편향되어야 하지만, 학습결과는 각 모델별로 서로 다른 결과를 도출한 것에 기인한다.

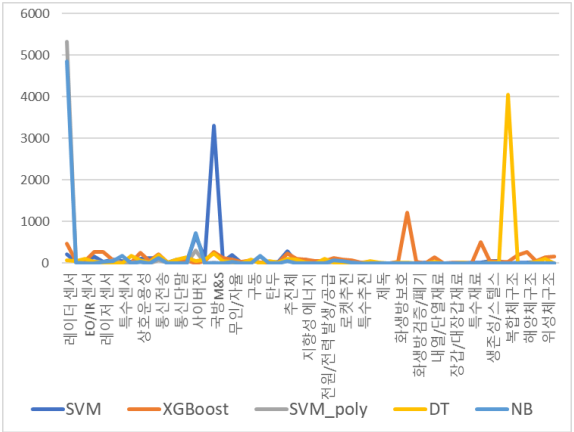

그림 2. 각 분류모델에서의 ADD 기술분류 그래프

중분류(Lv2)에서 확인한 분석내용을 토대로 보다 상세한 분석을 위해 그림 3 과 같이 소분류(Lv3)에 대한 기술분류를 통해 상세분석을 수행하였다. 중분류 결과에서 상대적으로 낮은 편향성을 보인 XGBoost 모델에 대해서 소분류(Lv3) 기술분류를 수행하였으며, 결과에서도 '제독제' 항목에 편향되는 결과를 보였다. 여기에서 유추가능한 모델의 특성은 XGBoost 의 경우, 학습데이터가 부족한 기술항목의 용어에 상대적으로 높은 가중치를 부여한 것으로 추정된다.

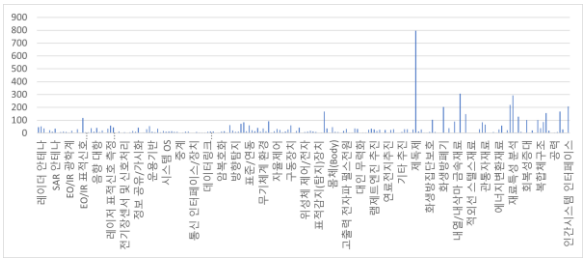

그림 3. XGBoost 모델의 소분류 예측 결과

학습 데이터셋 대비 편향된 결과를 도출하는 것에 대한 원인 분석의 의견은 다음과 같다.

첫째, 원시데이터셋의 기술분류 항목별 데이터 수량 편차로 인한 학습 편중의 쏠림 현상이다. 예를 들어, 국방과학기술 소분류 192 종 중 '레이더 송수신', '레이더 안테나', '통신 모뎀' 등과 같이 국방분야 외 타 과학기술분야에서도 활발한 연구가 진행되는 기술은 논문과 특허 실적이 많으나, '기폭장치', '전자재밍', '화학응용장치' 등과 같이 국방분야 또는 특수분야에서 한정적으로 연구 활용하는 기술은 특허 및 논문의 실적이 상대적으로 적다. Tf-Idf 로 데이터의 편중성을 낮추려 노력하였으나 통계기반의 학습모델은 각 모델의 특성에 따라 데이터의 과적합 경향을 완전히 제거하지는 못한 것으로 추정한다.

둘째, 자연어 전처리의 미흡이다. 기계학습 모델의 학습 전 반복적으로 수행한 불용어의 식별, 금지어 제거, 기술용어 및 전문용어 등록 등의 과정은 원시데이터셋을 본 의미를 제한하게 된다. 또한 식별하지 못한 단어는 학습 시 기계학습 모델의 특성에 따라 가중치가 서로 다르게 부여되어 기술분야별 핵심 용어와 상관없이 비중이 높게 책정된 것으로 추정된다.

셋째, 다중클래스 및 다중레이블 미적용에 따른 영향이다. 특허 및 논문은 하나의 카테고리로 명확하게 분류하기 어려운 경우가 많다. 기술융합의 시도처럼 하나의 기술문서 내에서도 여러 주제가 담길 수 있으나, 본 연구의 원시데이터 레이블링 작업에서는 가장 영향도가 높은 기술 카테고리 하나만 선택하였다. 이로 인해 분류모델의 성능에 영향을 미쳤을 것으로 추정한다.

4. 결론 및 향후 계획

본 연구결과는 3.5 절의 실 데이터 입력 분석 평가에서 서술한 바와 같이 실제 기관 및 기업의 기술실적을 분석으로 적용하기에는 어려움이 있다. 따라서,

미흡한 원인으로 추정되는 세 가지 항목에 대한 추가 연구는 학습계수 조정, 파라메터 최적화, 자연어 전처리를 위한 말뭉치 및 단어사전 보완, 다중클래스 및 다중레이블 반영 등으로 해결하고자 한다. 또한, 기계학습의 방식을 이해하고 인공지능 신경망 방식의 연구 수행을 통해 분류모델의 성능을 높이도록 수행할 계획이다.

참고문헌

[1] 홍유정, “융합 R&D 전략이 과학기술성과에 미치는 영향에 관한 연구: 대학의 정부 R&D 사업을 기반으로”, 한국혁신학회지, 제 16 권, 제 3 호, 31p, 2021
[2] 전상규, “특허 네트워크 분석을 통한 기술융합 및 융합기술의 확산 연구 – 디지털 데이터 처리 기술 중심으로”, 지식재산연구, 제 16 권, 제 4 호, 42p, 2021
[3] 김성훈, “특허문서 분류를 위한 딥러닝 개별 모델 분류기 성능비교”, 대한전자공학회 하계학술대회 논문집, 2021, 3P
[4] 심우철, “한국 특허문헌 특성 및 딥러닝 기반 분류 모델을 고려한 CPC 자동분류에 관한 연구”, 한국소프트웨어종합학술대회 논문집, 2020, 3p
[5] 이상우, “전자파 인체영향 연구논문에 대한 연구형태 자동분류 연구”, 한국전자파학회논문지, 제 31 권 제 10 호, 4p, 2020
[6] 한국과학기술기획평가원, “국가과학기술표준분류체계 개정 프로개스 개선 및 전면 개정을 위한 기획연구”, 과학기술정보통신부, 2018
[7] 석제범, “4 차 산업혁명을 선도하는 주요기술대상-기술수준평가 및 기술수준 향상방안”, 대전시, 정보통신기술진흥센터, 2021
[8] 국방기술표준분류, “https://dtims.dtaq.re.kr/vps/OINF_searchStndList.do”, DTiMS 열린정보마당, 2014