기계학습을 이용한 특허 분류의 성능 비교에 대한 연구

이주현O, 강지호*, 박아람*, 박상성**, 장동식*
O*고려대학교 산업경영공학과
**고려대학교 기술경영전문대학원
e-mail: leeju@korea.ac.krO, hanyul@korea.ac.kr**

Performance Comparison of Patent Classification Using Machine Learning Models

Juhyun LeeO, Jiho Kang*, Ahram Park*, Sangsung Park**, Dongsik Jang*
O*Dept. of Industrial Management Engineering, Korea University
**Graduate School of Management of Technology, Korea University

● 요 약 ●

최근 특허분석의 중요성이 부각되고 있다. 특허분석을 위해 검색된 특허 중 노이즈 특허를 분류하는 작업은 많은 시간과 비용을 요구한다. 본 논문에서는 효율적인 특허분석을 위한 노이즈 특허 분류 성능의 비교를 진행한다. 그리고 비교한 결과를 통해 노이즈 특허 분류에 최적의 모형을 찾는 것을 목표로 한다. 듀얼 카메라 특허 603건을 이용하여 실제 실험을 실시한 결과, 나이브 베이지안 분류 모형의 분류 성능이 가장 우수하였다.

키워드: 특허분석(patent analysis), 기계학습(machine learning), 텍스트 마이닝(text mining)

I. Introduction

특허분석은 제 4차 산업혁명과 빅데이터가 결합된 연구 분야이다. 제 4차 산업혁명이 시작되면서 많은 융합 기술이 개발되고 있으며, 대부분이 특허로 출원되고 있다. 이러한 관점에서 급속도로 성장하고 있는 기술시장을 따라가기 위해 특허분석은 필수적이다. 그러나 특허를 분석하기 위해서는 특허 데이터베이스로부터 특허를 검색하고, 그 중에서 노이즈 특허를 분류하는 작업이 선행되어야 한다. 그 이유는 Kang et al.(2012)들이 지적한 바와 같이, 특허 검색 시스템이 현재 기본적으로 불리언 모델을 기반으로 사용자가 질의한 키워드를 포함하는 검색 결과를 보여주기 때문이다. 따라서 효율적인 특허분석을 위해서 노이즈 특허를 분류하는 알고리즘의 개발이 필요하다.

문서분류를 위한 연구가 다양한 분야에서 진행되어 왔다. Diao et al.(2000)들은 이메일에 대해 개인의 흥미에 따른 필터링 알고리즘을 구성하는 연구를 진행하였다. Singh et al.(2014)들은 4가지 문서 전처리 과정과 7가지 기계학습 알고리즘을 조합하여 문서분류의 성능을 비교한 연구를 진행하였다. 그러나 특허는 서지사항, 요약, 청구범위 등이 포함된 복잡한 구조로 이루어져 있기 때문에 분류를 하는 것에 어려움이 따른다. 따라서 본 연구에서는 텍스트 마이닝 기법을 이용해 특허에서 사용되는 키워드를 중심으로 특허를 분류하는 최적의 모형을 탐색한다.

II. The Proposed Scheme

텍스트 마이닝 기법을 이용해 기계학습 모형을 학습하고 노이즈 특허를 효율적으로 분류하는 알고리즘을 찾기 위한 흐름도는 그림 1과 같다. 우선 특허 데이터베이스로부터 특허를 수집한다. 그리고 문서별 단어의 빈도를 고려하는 Document-Term 행렬(DTM)을 만든다.

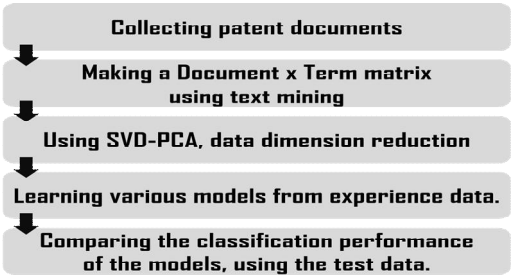

Fig 1. Flow chart of the proposed scheme

구성된 DTM은 차원이 상당히 크기 때문에 차원축소를 할 필요가 있다. 단어(Term)의 수는 문서(Document)의 수보다 훨씬 크기 때문에 SVD-PCA 방법을 사용하여 차원을 축소한다. 그 다음 원 데이터를 학습 데이터와 검증 데이터로 나누어 학습을 진행한 뒤, 노이즈 분류 성능을 비교한다.

III. Experiment and Results

노이즈 특허를 분류하기 위한 최적 모형을 찾기 위해 실제 특허문서를 사용하여 실험을 진행한다. 실험에 사용된 데이터는 WIPS ON 특허 데이터베이스로부터 검색한 듀얼 카메라 특허이다. 검색 결과, 미국 특허 603건을 얻을 수 있었다. 노이즈 특허분류의 성능을 확인하기 위해 로지스틱 회귀모형(LR, Logistic Regression), 의사결정나무모형(DT, Decision Tree), k-최근접 이웃모형(k-NN, k-Nearest Neighbor), 나이브 베이지안 분류 모형(NB, Naive Bayesian)을 사용하였다.

603건의 미국 듀얼 카메라 특허에 사용된 단어는 모두 2,735개였다. 문서의 수(n)와 단어의 수(p)의 관계가 n << p이므로 일반적인 통계모형에 적용하기 어렵다. 따라서 SVD-PCA를 사용하여 차원을 축소하였다. SVD-PCA를 적용한 결과, 40개의 주성분을 선택할 때, 원래 정보의 90%를 유지할 수 있었다. 따라서 본 논문에서는 40개의 주성분을 이용한다.

모형의 분류 성능을 비교하기 위해 603건의 특허 중 402건을 훈련 데이터로, 201건을 검증 데이터로 사용하였다. 각 모형을 학습한 뒤, 분류 성능을 검증한 결과는 표 1과 같다.

Table 1. Result of Performance Comparison

Value	LR	DT	k-NN	NB
Acc	0.69	0.84	0.91	**0.87**
Sen	**0.79**	0.11	0.00	0.37
Spe	0.68	**0.92**	1.00	**0.92**
Pre	0.21	0.12	None	**0.33**
F1	0.33	0.11	None	**0.35**

학습 데이터와 검증 데이터로 최적의 모형을 찾는 작업을 실시하였다. LR의 경우, Stepwise 변수선택 절차를 이용하였으며 DT는 가지치기 작업을 실시하였다. k-NN의 경우, k=7인 경우가 가장 최적이었으나 모든 특허를 노이즈라고 예측하여 분류의 의미가 없었다. 모형의 분류 성능을 측정하기 위해 정확성(Acc, Accuracy), 민감도(Sen, Sensitivity), 특이도(Spe, Specificity), 정밀도(Pre, Precision), F-1 점수(F1)를 사용하였다. 4가지 모형에 대하여 5가지 분류 성능 지표를 교차 비교한 결과, NB가 5개 중 4개의 지표에서 가장 우수한 성능을 보였다.

IV. Conclusion

효율적인 특허분석을 위해 노이즈 특허를 잘 분류하는 모형을 탐색하는 연구를 진행하였다. 미국 듀얼 카메라 특허 603건으로 실험을 진행한 결과, 선행 연구에서 문서 분류에서 강점을 보였던 나이브 베이지안 분류 모형의 성능이 가장 우수함을 알 수 있었다.

향후 연구에서는 노이즈 특허 여부가 보다 객관적으로 라벨링된 원데이터를 이용한 연구가 필요할 것으로 예상된다.

ACKNOWLEDGEMENT

본 논문은 2015년 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임.(한국연구재단-NRF-2015R1D1A1A01059742). 본 논문은 2017년 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임.(한국연구재단-NRF-2017R1A2B1010208). 본 논문은 BK21 플러스사업(고려대학교 제조·물류 분야에서의 빅데이터 운용 사업팀)으로 지원된 연구임.

REFERENCES

[1] Y. Diao, H. Lu and D. Wu, “A comparative study of classification based personal e-mail filtering”, Knowledge Discovery and Data Mining, Vol. 180, pp. 408-419, Apr. 2000.

[2] M. Kang, J. Song and W. Lee, “Searching Patents Effectively in terms of Keyword Distributions”, Journal of Information Technology and Architecture, Vol. 9, No. 3, pp. 323-331, Sep. 2012.

[3] P. Singh, S. Verma and O. P. Vyas, “Software fault prediction at design phase”, Journal of Electrical Engineering & Technology, Vol. 9, No. 5, pp. 1739-1745, Sep. 2014.

[4] Y. Kim, J. Lee, J. Lee, J. Kang, S. Park and D. Jang, “Comparison of Classification Performance of PCA and LDA Using Patent Document”, Proceedings of KIIS Spring Conference, Vol. 27, No. 1, pp. 89-90, Apr. 2017.