Doc2Vec 을 이용한 특허 문서 자동 분류

송진주, 강승식
국민대학교 소프트웨어융합대학
e-mail: sskang@kookmin.ac.kr

Automatic Classification of Patent Documents Using Doc2Vec

Jinjoo Song and Seung-Shik Kang
School of Software Technology, Kookmin University

요 약

지식과 정보의 중요성이 강조되는 지식기반사회에서는 지식재산권의 대표적인 유형인 특허의 중요성이 날로 높아지고 있고, 그 수 또한 급증하고 있다. 특허 문서의 효과적 검색과 이용을 위해서는 새롭게 출원되는 특허 문서의 체계적인 분류 작업이 선행되어야 하고, 따라서 방대한 양의 특허 문서를 자동으로 분류해주는 시스템이 필요하다. 본 연구에서는 Doc2Vec 모델을 이용하여 국내 특허 문서의 특징(feature)을 추출하고, 추출된 특징을 바탕으로 한 특허 문서의 자동 분류 모형을 제안한다. 먼저 국내에 등록된 31,495 건의 특허 문서의 IPC(International Patent Classification)와 요약 정보를 바탕으로 Doc2Vec 모델을 구축하였다. 구축된 Doc2Vec 모델을 통하여 훈련데이터의 특징을 추출한 후, 이 특징 벡터를 이용하여 분류기를 학습하였다. 마지막으로 Doc2Vec 모델을 이용하여 실험데이터의 특징 벡터를 추출하고 분류기의 성능을 실험한 결과, 43%의 분류 정확도를 얻었다. 이를 통해, 특허 문서 분류 문제에 Doc2Vec 모델의 사용 가능성을 확인할 수 있었다.

주제어: 특허 문서 분류, IPC 자동 분류, 문장 임베딩, Doc2Vec

1. 서론

오늘날과 같이 지식과 정보의 중요성이 강조되는 지식기반사회에서 지식재산권의 대표적 유형인 특허의 중요성도 날로 증대되고 있다. 그 결과 매년 출현되는 특허의 수 또한 급증하고 있다. 특히, 우리 나라의 경우 특허 출원 강국으로, 2016 년의 경우, 중국(134 만건), 미국(60.6 만건), 일본(31.8 만건)에 이어 우리 나라가 20.9 만건으로 4 위를 차지하였고, GDP 및 인구 대비 특허 출원 건수에서는 세계 1 위를 차지하였다[1]. 이러한 상황 하에, 새롭게 특허를 출원하는 경우와 기존의 특허를 효과적으로 이용하기 위해서는 빠르고 용이한 특허 검색 시스템이 필요하고, 이를 위해서는 특허 문서의 체계적인 분류 시스템이 선행되어야 한다. 현재는 전문인력이 일일이 특허의 내용을 파악하여 기술적 주제에 따라 분류함에 따라 많은 시간과 비용이 소요되고 있다.

따라서 특허 문서 자동 분류 시스템에 대한 요구와 관심이 높아지고 있다. 기존 연구로는 문서 분류에 적합한 특징을 추출하는 연구들[2-7]이 주를 이루고 있고, 문서 분류 기법에 대한 연구[8]도 있다. 최근에는 딥러닝을 이용한 특허 문서 분류 연구[9]도 등장하였다. 그러나 기존 문서 분류와 달리, 특허 문서 분류는 분류 기준이 세분화되어 있어, 다른 분류코드를 갖는 문서들 간에도 상당 부분 동일한 키워드가 존재하는 경우가 많고, 이는 분류 정확도를 낮추는 요인으로 작용하고 있다.

본 논문에서는 특정 도메인을 대상으로 Doc2Vec 모델을 구축하고, 이렇게 구축된 Doc2Vec 모델을 이용하여 특허 문서의 특징 벡터를 구성하는 방식으로 특허 문서 분류 성능 향상을 꾀하고자 한다. 문서 분류기 학습에는 SVM(support vector machine)과 로지스틱 회귀(Logistic Regression) 기법을 이용하였다.

본 논문의 구성은 다음과 같다. 2 장에서 관련 연구를 리뷰하고, 3 장에서 Doc2Vec 을 이용한 특허 문서 분류에 대해 설명하고, 4 장에서는 실험결과를 분석한다. 마지막 장에서 결론 및 향후 연구에 대한 논의로 논문을 마무리한다.

2. 관련 연구

2.1 IPC 구조 및 특허 문서 구조

특허 문서에 대해 전세계적으로 통일된 분류와 검색을 위해 1954 년에 국제특허분류(International Patent Classification, IPC)가 만들어졌다. IPC 의 구성은 섹션(8 개), 클래스(128 개), 서브클래스(약 650 개), 메인그룹(약 6,800 개), 서브그룹(65,000 개 이상)으로 이루어진다. 이 중 섹션과 서브섹션은 알파벳으로 표시하고 나머지는 숫자로 표시한다. 메인그룹과 서브그룹 사이에는 ‘/’ 기호를 써서 구분한다. (예, A63F13/352) 하나의 특허 문서는 해당 기술에 따라 하나 이상의 IPC 를 갖는다.

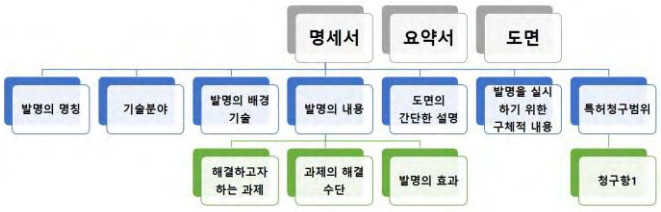

그림 1. 특허 문서의 구조

특허 문서의 구조는 그림 1 과 같다. 크게 '명세서', '요약서', '도면'으로 구성되고, '명세서'는 다시 발명의 내용을 설명하는 부분과 특허 청구 범위로 나뉜다. 본 연구에서는 IPC 와 요약 정보를 각각 클래스 레이블과 문서로 사용하여 Doc2Vec 모델을 구축하였다.

2.2 특허 문서 분류 연구

문서 분류기의 성능을 좌우하는 가장 핵심적인 요인은 문서를 대표하는 특징 추출 방법이다. 따라서 특허 문서를 대표하는 특징으로 무엇을 사용할지에 대한 연구가 기존 연구의 대부분을 차지하고 있다. 추출된 특징으로 벡터를 구성하기 위해 다양한 방법을 사용하였는데, 문서의 의미적 구조정보를 이용한 특허 문서 분류[2], 사용 품사에 따른 분류 정확도[3] 연구 등이 있다. 또한 박찬정 외[5]는 카이제곱 통계량, 정보이득, 상호정보량, 우세정보량을 이용하여 각각의 특징 벡터를 구성하고, 이를 KNN(K-Nearest Neighbor)을 이용하여 특허 문서를 분류하였다.

특허 문서를 구성하고 있는 여러 항목의 사용 유무에 따른 분류 성능 연구도 진행되었다. Larkey[7]는 제목, 요약, 배경, 요약의 일부, 청구항을 사용하여 미국 특허 문서를 분류하였고, 강민규 외[3]는 문서 전체, 제목, 요약, 청구항, 요약-청구항을 사용한 경우의 한국 특허 문서 분류 성능을 비교하였다. 임소라 외[6]는 배경기술과 기술분야 항목을 사용하여 특허 문서를 분류하였다.

Chen 외[10]는 IPC 서브그룹 레벨에서 SVM 과 KNN 을 계층적으로 이용하여 문서 분류를 시도하였다.

3. Doc2Vec 을 이용한 특허 문서 자동 분류

3.1 Doc2Vec 모델 구축

문장 임베딩의 한 방법인 Doc2Vec 은 BoW (bag-of-words) 기반의 특징 추출 방식의 한계-단어의 순서 정보 상실과 단어의 의미론적 정보 무시-를 극복하고자 제안되었다. Doc2Vec 은 문서를 수치화한 벡터로 표현하는 것을 목표로 한다.

기존의 Word2Vec 의 확장 형태인 Doc2Vec 에는 분산 메모리와 분산 BoW 의 2 가지 유형이 있다[11]. 그림 2 에서처럼, Word2Vec 의 CBoW (continuous bag-of-words) 모델에 문서 ID 벡터를 추가한 것이 분산 메모리 유형이고, 분산 BoW 유형은 Skip-gram 모델의 단어 정보 대신 문서 ID 를 부가한 것이다.

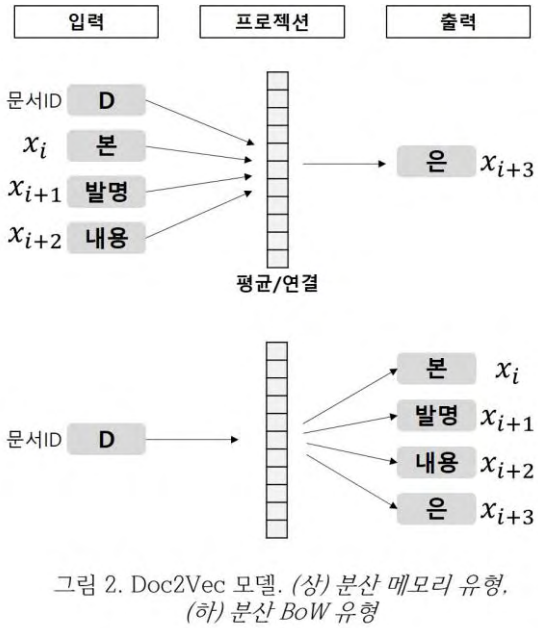

그림 2. Doc2Vec 모델. *(상) 분산 메모리 유형, (하) 분산 BoW 유형*

3.2 문서 분류기 학습

본 연구에 사용된 분류 기법은 SVM 과 로지스틱 회귀이다. SVM 은 데이터를 여러 개의 클래스로 분류할 때, 클래스 간의 거리인 마진을 최대화하는 분류 경계면을 찾는 기법이다. 로지스틱 회귀는 범주형 변수를 예측하는 모델로 SVM 과 마찬가지로 분류기 모형 구축에 널리 사용되고 있다.

4. 실험

4.1 데이터셋 및 실험 환경

본 논문에서 분류에 사용된 데이터셋은 '게임 기술'에 관한 데이터셋[3]이다. 따라서 '게임 기술'에 관한 특허 문서를 바탕으로 Doc2Vec 모델이 구축되었다.

먼저 Doc2Vec 모델 구축을 위해, 특허정보넷 키프리스 에서 '게임 기술'을 검색어로 하여 2000 년에서 2018 년까지 등록된 31,495 건의 특허 문서를 다운로드하였다. 특허 문서의 여러 정보 중, IPC 정보를 클래스 레이블로, 요약 정보를 문서로 사용하여 Doc2Vec 모델을 구축하였다.

분류기 학습에 사용된 데이터는 강민규 외[3]의 특허 문서 데이터셋을 이용하였다. 해당 데이터셋은 '게임 기술'과 관련된 국내 특허 문서 1,229 건으로 구성되어 있다. 학습데이터는 296 건, 실험데이터는 933 건이다. 자세한 데이터셋 정보는 표 1 에 나타나 있다. 데이터셋의 특허 문서 항목 중, 실험에 사용된 항목은 요약과 청구항이다.

분류 실험은 Intel coreTM i5-3470 CPU, 12GB RAM 의 데스크탑의 64bit Linux OS 상에서 수행되었고, Doc2Vec 모델은 Gensim 라이브러리를, SVM 과 로지스틱 회기 분류기는 scikit-learn 라이브러리를 이용하여 Python3.5 환경에서 구현하였다. 모델 훈련에 사용된 Doc2Vec 알고리즘은 분산 BoW 유형이고, 이 때, 단어 윈도우 사이즈는 8, learning rate 는 0.025 를 사용

하였다.

<표 1> 강민규 외[3] 특허 문서 데이터셋

구분	문서	클래스	클래스 내 문서
학습데이터	296 건	19 개	7~20 건
실험데이터	933 건	15 개	7~304 건

표 2 는 Doc2Vec 모델을 통해 추출한 차원 수에 따른 각 분류기의 분류 정확도를 나타낸다. 가장 높은 분류 정확도는 550 차원의 특징 벡터와 로지스틱 회귀를 사용하였을 때로, 43.00%의 정확도를 나타내고 있다. 비선형 SVM 에서는 250 차원 특징 벡터 사용 시, 가장 높은 성능을 보여주고 있고, 선형 SVM 과 마찬가지로 선형인 로지스틱 회귀에서는 550 차원에서 가장 정확도가 높았다. 이는 19 개의 클래스를 분류하는 분류기 모형 학습이라는 복잡한 문제를 풀기 위해 선형 분류기에서는 비선형 분류기 보다 더 많은 특징을 필요로 하기 때문인 것으로 보인다.

<표 2> 제안방법의 특허 문서 분류 정확도 (%)

차원	비선형 SVM	선형 SVM	로지스틱 회귀
100	31.30	27.01	31.83
150	32.69	28.51	32.26
200	35.26	32.80	33.33
250	37.51	35.16	36.01
300	32.48	34.41	35.69
350	30.01	38.59	36.87
400	30.55	37.83	39.44
450	27.97	38.59	40.41
500	24.97	37.94	41.69
550	26.37	42.23	**43.09**
600	24.54	36.98	40.73

표 3 은 같은 학습/훈련 데이터셋을 사용한 기존 방법과 제안 방법의 분류 정확도 비교를 나타내고 있다. 강민규 외[3] 방법에서는 TF-IDF 로 특징을 추출하고 SVM 을 통하여 분류 실험을 하였는데, 분류 정확도가 49.03 으로 제안 방법보다 높다. 그러나 제안 방법도 43.09 의 정확도를 보여, Doc2Vec 모델을 사용한 특허 문서 분류 방법의 가능성을 보였다. 제안 방법의 분류 정확도가 낮은 이유는 19 개의 클래스를 구분하는 문제에 적용하기에는 Doc2Vec 모델 구축에 사용된 특허 문서의 양과 수가 적기 때문일 것으로 추정된다. 따라서 Doc2Vec 모델 구축에 더 많은 특허 문서의 항목을 사용하고, 그 수를 늘려서 추가 실험을 해볼 필요가 있다.

<표 3> 기존 방법과의 비교 (%)

구분	강민규 외[3]	제안 방법
분류 정확도	**49.03**	43.09

5. 결론

본 논문에서는 Doc2Vec 을 이용한 특허 문서 분류 방법을 제안하였고, 실험을 통하여 Doc2Vec 모델을 이용한 문서 분류의 가능성을 보였다. 특정 도메인에 특화된 Doc2Vec 모델을 만들고 이를 이용하여 특징을 추출한 후 SVM 과 로지스틱 회귀 기법으로 분류기를 학습하였다. 실험 결과, 로지스틱 회귀를 이용한 방법이 SVM 을 이용한 방법보다 조금 더 좋은 분류 성능을 보였다. 제안 방법의 경우, 기존의 방법보다 분류 정확도가 낮았는데, 이는 Doc2Vec 모델 구축에 사용된 특허 문서 수가 충분하지 않았기 때문으로 추정된다.

향후 과제로 더 많은 특허 문서를 수집하여 Doc2Vec 모델을 구축할 필요가 있다. 또한 실제로는 특허 문서 분류에서 다중 IPC 를 부여하고 있기 때문에, 단일 IPC 분류에서 다중 IPC 분류로 확장된 연구 수행이 필요하다.

감사의글

이 논문은 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학지원사업(2016-0-00021)과 한국연구재단의 중견연구지원사업(2017R1A2B4011015)의 연구결과로 수행되었습니다.

참고문헌

[1] “World Intellectual Property Indicators 2017”. (2017). WIPO.

[2] 김재호, & 최기선. (2005). 문서의 의미적 구조정보를 이용한 특허 문서 분류. *한국정보과학회언어공학연구회 학술발표논문집*, 28-34.

[3] 강민규, 정인상, & 강승식. (2009). 특허 문서 분류를 위한 영역-자질 선택 방법. *한국정보과학회 학술발표논문집*, 36(1C), 284-287.

[4] 박찬정, 김기용, 성동수, & 이건배. (2013). KNN 알고리즘을 이용한 특허문서의 다중 IPC 자동 분류. *한국정보과학회 학술발표논문집*, 1502-1504.

[5] 박찬정, 김기용, & 성동수. (2014). KNN 을 이용한 융합기술 특허문서의 자동 IPC 분류. *한국정보기술학회논문지*, 12(3), 175-185.

[6] 임소라, & 권용진. (2017). 특허문서 필드의 기능적 특성을 활용한 IPC 다중 레이블 분류. *인터넷정보학회지*, 18(1), 77-88.

[7] Larkey, L. S. (1999). A Patent Search and Classification System. In *Proc. ACM Conf. Digital Libraries*. 179-187. ACM.

[8] 강지호, 김종찬, 이준혁, 박상성, & 장동식. (2016). 특허 문서 분류 알고리즘 비교 연구. *한국지능시스템학회 학술발표논문집*, 26(1), 9-10.

[9] 장시온, 서원철, & 최성철. (2017). 딥러닝 기반의 특허문서 다중분류. *한국경영과학회 학술대회논문집*, 5554-5566.

[10] Chen, Y. L., & Chang, Y. C. (2012). A Three-Phase Method for Patent Classification. *Information Processing & Management*, 48(6), 1017-1030.

[11] Le, Q., & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. In *Int. Conf. Machine Learning*. 1188-1196.