Information Systems Review
Vol. 22, No. 1 February 2020 http://dx.doi.org/10.14329/isr.2020.22.1.017

LDA 토픽 모델링과 Word2vec을 활용한 유사 특허문서 추천연구

LDA Topic Modeling and Recommendation of Similar Patent Document Using Word2vec

이 앞 길 (Apgil Lee) 한밭대학교 경영학과 박사과정
최 근 호 (Keunho Choi) 한밭대학교 경영회계학과 조교수
김 건 우 (Gunwoo Kim) 한밭대학교 경영회계학과 부교수, 교신저자

요 약

4차 산업혁명 시대의 시작과 함께 다양한 분야의 기술들이 서로 융합하며 새로운 형태의 기술과 제품들이 개발되고 있으며, 이와 더불어 그것들에 대한 시장 지배력을 갖기 위한 지식 재산권의 행사나 특허등록의 중요성이 높아지고 있어 국내는 물론 해외에서의 특허출원이 증가하고 있다. 이에 따라, 심사관 1인당 처리해야 할 특허 처리 건수가 해마다 많아지고 있어 선행기술조사에 소비되는 시간과 비용이 점점 증가하고 있는 실정이다. 본 연구는 다수의 해외특허 우선권 주장 시 동일 우선권 주장 특허문서 간 유사도를 계산하여 심사관 및 특허 출원인이 유사문서를 우선 검토 할 수 있도록 함으로써 심사 시간과 비용을 줄이고자 하였다. 이를 위해, 본 연구에서는 비정형 특허 문서의 데이터를 전처리 후 LDA 토픽 모델링과 Word2vec을 활용하여 특허 문서 간 유사도를 구하고, 이 유사도 점수가 높은 순으로 검토 문서를 우선 추천하는 유사 특허 추천 모델을 제안하였다. 3단계의 모델 생성과정을 통해 만들어진 모델을 사용하여 재현율 95%로 높은 결과를 보였다. 본 연구에서 제안한 모델을 통해, 심사관은 효율적으로 선행기술에 대한 조사가 가능해지며, 심사 수행 중 유사하다고 판단된 특허문서에 대한 심사 이력을 신속하게 참고할 수 있어 업무 부담감을 줄이고 심사품질을 향상시킬 수 있을 것으로 기대된다.

키워드 : 특허문서, 유사 특허문서 추천, LDA 토픽모델링, Text mining, Word2vec

Ⅰ. 서 론

1.1 연구의 배경

최근 들어 정보기술을 이용한 다양한 분야의 기술들이 자체적으로 서로 융합되어 새로운 형태의 제품과 기술들이 개발되고 있으며, 이와 더불어 그것들에 대한 지식 재산권의 일환으로 특허등록의 중요성이 높아지고 있어 국내는 물론 해외에서도 특허출원이 증가하고 있다(특허청, 2018).

특허는 권리를 인정받고자 하는 모든 국가에 출원하여 등록을 받아야 하며, 등록 받은 각 국가의

† 이 논문은 2019년도 한밭대학교 교내학술연구비의 지원을 받았음.

영토 내에서만 특허권 효력이 있다(이광희 등, 2015).

예를 들어 일본에서 특허가 등록되어 있다고 하더라도 한국에 등록이 되어 있지 않으면 한국에서는 특허 권리를 인정할 수 없다. 그로 인해 각 국가에서 권리를 행사하기 위한 해외특허출원이 증가하고 있다. 특허출원이란 특허를 받을 수 있는 권리를 가질 수 있는 자가 그 발명의 공개를 전재로 하여 특허청에 대하여 특허를 받고자 하는 의사를 객관적으로 정해진 양식에 맞게 표시하는 행위를 말한다. 특허 출원은 권리를 갖기 위해 해당 특허청에 문서를 작성하여 특허등록을 위해 제출하는 것이며 이때 작성한 서류를 출원서라 한다(이처영, 2001).

특허등록을 위해 작성한 정보 및 각 특허등록의 단계에 사용되는 모든 특허문서 정보를 특허정보라 하는데, 이 특허정보에는 출원된 기술의 내용 및 권리로 주장된 사항은 물론, 출원인 및 발명자의 인적 사항과 기타 서지사항 등에 대한 모든 정보가 포함된다. 본 논문에서는 특허등록을 위해 작성되고 심사 되어진 모든 문서를 특허문서로 통일하여 표현한다.

정보통신의 발달로 세계 교류가 활발해 짐에 따라 각 국가에서 권리를 행사하기 위한 해외특허 출원도 증가하고 있다. 국내에서 해외로 출원하는 건수는 2016년 기준으로 연간 미국으로는 약 3만 8,000건, 유럽으로는 6,500건이며 국내로 출원되는 건수는 4만 5,000건으로 매년 증가하고 있다(특허청, 2018). 이로 인해 심사관 1인당 처리해야 할 특허출원 처리건수가 많아지고, 특허 심사를 위해 조사해야 하는 선행기술 조사대상 특허문서의 양도 증가하기 때문에 심사에 드는 시간과 비용도 점점 증가하고 있는 실정이다. 미국의 경우, 심사관 1인당 1년에 처리하는 심사건수가 77건으로 1건당 약 26시간, 유럽의 경우 1인당 58건으로 1건당 약 34.5시간을 투입하는데 국내의 경우 1인당 217건으로 1건당 약 11시간으로 심사 투입시간이 선진국의 절반에도 못 미치고 있는 실정이다(특허청, 2018). 적은 심사 투입시간은 특허 품질 저하를 초래할 수 있고 '부실특허'로 이어질 수 있다. 이와 같은 문제는 심사인력의 증원만으로는 어렵기 때문에 특허를 출원한 기술에 대한 심사 기간과 비용을 줄이기 위해 많은 연구가 진행되어 왔다.

강지호 등(2017)의 연구에 따르면 일반적으로 특허 출원 심사를 위해 기존 특허 문서를 탐색하는 절차는 출원서의 발명의 명칭(Title), 요약(Abstract), 청구항(Claims) 등의 내용을 검토하고 권리범위를 구성하는 핵심 키워드를 파악한 뒤 이와 유사한 의미로 사용되는 확장 키워드를 파악하기 위해 기존 특허문서 검색을 반복적으로 수행한다. 또한 심사 대상 특허가 다수의 해외 우선권을 주장하는 특허의 경우에는 해외 각 나라의 심사정보나 같은 우선권을 주장하는 특허들을 참고하기 위해 패밀리조사를 하는데 그 패밀리 특허문서의 수가 무수히 많을 경우도 존재한다. 또한 패밀리 특허 간에는 다수의 기술이 복합적으로 있는 경우 검색 대상 특허와 상이한 기술의 특허도 포함되는 경우가 있어 하나하나 비교 검토하기에는 어려움이 있다. 이에 따라 해외 우선권 특허 출원 심사에 대해 심사관뿐만 아니라 특허 출원인의 경우도 선행기술조사에 많은 어려움을 겪고 있다. 따라서 본 연구는 해외 우선권 특허출원 심사 시 검색 대상 특허의 패밀리특허 간 유사도를 계산하여 유사도가 높은 특허문서를 추천하고자 한다. 특허문서에 관련된 대표 용어에 대한 정의는 제Ⅱ장 선행연구에서 설명한다.

1.2 연구의 목적

특허문서는 범위도 넓고 방대하며 다양한 정보를 담고 있다. 또한, 그 내용도 방대하기 때문에 전체 문서에서의 유사도 비교는 한계가 있다. 따라서 본 연구는 연구의 범위를 특허문서 전체로 하지 않고, 국가별 동일 용어의 표기방법이 다르고 그 범위도 넓어 심사의 어려움이 있는 해외 우선권 특허 출원 심사에서 사용되는 특허문서의 패밀리특허문서로 연구의 범위를 한정하고 해당 검색

대상 특허와 패밀리특허 문서 간의 유사도를 계산하기 위해 Text Mining 기법을 활용하고자 한다. 즉, 특허문서의 명칭, 초록, 청구항 각 부분의 비정형데이터를 수집하여 전처리한 후 LDA(Latent Dirichlet Allocation) 토픽 모델링과 Word2vec 알고리즘을 적용한다. 이 후, 실험을 통하여 특허문서 간 유사도 비교에 가장 효과적인 방법을 도출해 내며, 최종적으로 유사 특허 문서를 추천하는 모델을 제안하고자 한다. 이를 통해, 패밀리 특허조사를 필요로 하는 심사관은 업무 부담을 줄일 수 있고, 출원인은 효율적인 검색이 가능해짐으로써 심사품질 제고에 도움을 줄 수 있을 것으로 기대된다.

1.3 연구의 범위

특허문서의 유사성을 비교하기 위해서는 텍스트 데이터를 수치적인 방식으로 표현하고 그 수치에 언어의 의미를 반영해야 하는데, 텍스트 데이터에는 이는 시대적 상황, 민족적, 언어적, 그리고 문화적 시대적 상황, 민족적, 언어적, 그리고 문화적 요소가 담겨 있기 때문에 텍스트를 이해하고 분석하는 것은 매우 어려운 작업이다. Hearst(1999)는 5가지 이유로 텍스트 데이터를 다루는 것이 어렵다고 설명하고 있다.

첫째, 추상적인 개념을 표현함에 있어서의 모호하다. 특정 한 단어로 표현 되는데 한계가 있다. 둘째, 개념들간에 미묘하고 추상적인 관계의 수많은 조합이 존재한다. 개념과 개념 사이의 관계는 시대나 문화, 사회적 상황, 문맥에 따라 여러 가지로 규정될 수 있다. 셋째, 유사한 개념(동의어, 유의어)을 표현하기 위한 다양한 방법이 존재한다. 한 개념에 대한다양한 표현방법은 텍스트를 효과적으로 분석하는 데 큰 어려움을 준다. 예를 들면 사람이라는 개념은 남자, 여자, 아이 등으로 표현할 수 있다. 또한 글자의 음은 같으나 뜻이 다른 낱말을 의미하는 동음이의어의 문제도 텍스트 분석을 어렵게 한다. 예를 들면 애플이라는 단어는 사과를 의미할 수도 있고 회사를 의미할 수도 있다. 넷째, 텍스트의 고차원성. 텍스트 데이터는 각각의 고유한 단어를 하나의 차원(dimension) 또는 자질(feature)로 하기 때문에 대부분이 고차원이다. 단어의 조합으로 된 다차원적인 문서집단은 텍스틀 분석하는데 있어 어려움을 준다. 다섯째, 개념을 시각화하기 어렵다. 단어로 표현되는 개념의 다차원성으로 인해 개념의 시각화가 용이하지 않다.

이러한 텍스트 분석의 어려움을 해결하기 위해 많은 연구가 진행되고 있는데 본 연구에서는 특허문서의 특성에 맞는 효과적인 모델을 찾기 위해 3단계로 구성된 실험을 수행하였다.

첫 번째 단계는 Word embedding의 대표적인 방법인 Word2vec과 토픽 모델링의 대표적인 방법인 LDA 그리고 Word2vec과 LDA을 통합한 방법들 중 우수한 방법을 도출하는 단계이며, 두 번째 단계는 도출된 방법의 최적 Parameter 값을 찾는 단계이다. 세 번째 단계는 앞선 단계의 방법을 적용하여 유사도 계산 시 특허문서의 명칭, 초록, 청구항별 최적 가중치를 찾고 최종적으로 생성된 모델을 사용하여 패밀리 특허문서의 유사문서를 추천하는 단계이다.

본 연구의 구성은 다음과 같다. 제Ⅱ장에서는 텍스트 데이터를 분석하기 위한 선행연구들과 그 방법들에 대해 살펴보고, 제Ⅲ장에서는 모델 생성을 위한 3단계에 대해 자세히 다룬다. 제Ⅳ장에서는 최종 선택된 모델의 성능에 대한 실험을 수행하고, 그 결과를 제시한다. 마지막으로 제Ⅴ장에서는 결론 및 향후 연구에 대하여 논의한다.

Ⅱ. 관련 연구

특허문서는 다양한 정보를 담고 있으며 그 내용도 공신력 있기 때문에 다양한 분야의 연구가 수행되어 왔으며 현재도 많은 연구들이 진행되고 있다. 또한 그 범위도 넓고 방대하여 빅데이터 기반 기술을 이용한 연구도 함께 진행되고 있다. 이원상 등(2013)은 지속적으로 축적한 삼극 특허 데이터의

IPC(International Patent Cooperation) 코드를 빅데이터 기술을 적용하여 기술 융복합 패턴을 예측하는 연구를 수행하였고, 김갑조 등(2017)은 특허문서의 텍스트 데이터를 수집하여 토픽 모델링을 적용해 유망기술을 예측하는 연구를 수행하였다. 고광수 등(2011)은 텍스트 마이닝을 이용하여 특허 문서의 전체 텍스트를 TF-IDF를 활용하여 단어의 가중치로 특징치를 찾아 유사한 문서 순으로 우선적으로 배치하여 검색에 효율을 높일 수 있는 방법을 연구하였다. 하지만 특허문서는 명칭, 초록과 같이 함축적인 표현을 하는 부분과, 청구항과 같은 세부사항을 상세하게 기술하는 부분의 유형별로 각각의 단어가 다른 의미와 정도를 가지기 때문에 한 부분만 사용하거나 전체를 통합하여 사용하게 되면 특허의 특질을 제대로 반영하지 못하는 문제가 발생할 수 있다.

2.1 특허

특허란 발명자에게 기술을 공개한 대가로 일정 기간의 독점권을 부여하는 제도로 특허권 확보를 위한 출원에서 등록까지의 많은 절차를 거쳐야 한다. 일반적인 서지사항은 출원번호, 출원일자, 공개번호 등의 출원 데이터와 출원인에 관한 인적사항 등이 있으며 출원 문서의 구성을 자세히 살펴보면 다음과 같다.

2.1.1 발명의 명칭

발명의 카테고리가 명료하게 구분될 수 있도록 출원을 간단명료하게 압축된 의미로 기재하는 부분으로 출원의 기술적인 분류, 정리 및 조사에 쉽게 활용하기 위해 표현된 명칭이다. 언어를 표현하는 방식은 각 국가마다 다르기 때문에 같은 우선권을 주장하는 특허문서도 각 국가별로 발명의 명칭이 다를 수 있다.

2.1.2 초록

특허의 상세한 설명을 요약한 것으로 중요한 키워드 중심으로 작성된 부분이다.

2.1.3 청구항

특허 권리를 보호받고자 하는 발명 기술의 요지와 권리범위를 작성한 부분으로 특허 심사 시 신규성, 진보성, 선원주의 등을 판단하는 기준이자 특허등록 후 권리 행사 시 권리범위 판단의 기준이 되는 부분이다. 범위를 보면 첫째 발명의 상세한 설명에 의여 뒷받침 될 것, 둘째 명확하고 간결하게 작성할 것, 셋째 발명의 구성에 없어서는 아니 되는 사항으로만 기재될 것, 따라서 상세한 설명에 기재되지 않은 내용을 청구한 경우, 필요이상으로 장황하게 기재하여 발명의 구성을 불명확하게 하는 경우, 발명의 구성요소를 단순히 나열만 하여 각 구성 요소 간의 상호작용을 파악할 수 없는 경우 등은 허용되지 않는다(이처영, 2001).

2.1.4 발명의 상세한 설명

발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 사용 할 수 있을 정도로 그 발명의 목적, 구성, 효과가 기재된 부분으로 청구범위에 기재된 용어를 명확하게 정의함과 동시에 청구범위를 해설하는 역할을 하는 것으로 그 기재가 모순되어서는 안 된다(이처영, 2001).

이 밖에도 도면, 명세서 등 많은 정보가 있지만 본 연구에서 사용하는 범위는 출원 심사를 위해 기존 특허 문서의 탐색 절차에서 내용 검토 및 권리범위를 구성하는 핵심을 파악하는 부분인 발명의 명칭(Title), 초록(Abstract), 청구항(Claims)으로 한정한다(강지호 등, 2017). 이는 함축적 유의어가 올 수 있는 부분으로 각 부분의 특성을 테스트를 통해 찾고 특성에 맞춰 모델에 적용한다.

2.1.5 국제특허출원

서론에서 설명한 내용과 같이 각 국가에서 권리행사를 하려면 각 국가에 특허등록을 해야 한다 그 방법은 다음과 같다.

2.1.5.1 파리조약에 기초한 개별국 출원

국내출원 후 1년 이내에 원하는 국가에 우선권 제도를 이용해 개별적으로 출원하는 방식으로써 우선권 제도란 파리조약의 동맹국의 제1국에서 정규출원을 한 자 또는 승계자가 1년 내에(우선권 기간) 다른 동맹국에 동일한 특허를 출원하여 우선권을 주장하는 경우 특허요건 등을 판단할 때 제1국의 출원일을 인정해주는 제도로 파리 조약의 우선권 제도는 언어 및 절차가 상이한 각 국가별로 모두 출원을 해야 하는 문제점을 해결하고자 도입되었다. 이와 같이 자국출원을 기초로 하여 해외 여러 나라에 출원하는 경우, 자국 출원과 관련된 모든 특허 및 출원을 패밀리 특허라 하는데, 패밀리 특허는 동일한 우선권(Priority)데이터를 가지며. 이것에 기초하여 각 국가에 출원한 내용과 각 국가에서의 출원에 대한 진행 사항에 대한 조사가 가능하다(이처영, 2001).

2.1.5.2 국제특허출원(국제협력조약 PCT에 기초한 출원)

국제협력조약(PCT)은 하나의 발명을 다수 국가에 출원을 하는 경우 그 출원을 용이하게 함으로써 출원인의 노력과 비용을 경감시키고, 각국 특허청의 심사부담도 경감시키고자 하는 조약을 말한다. 국제특허출원은 국제협력조약에 기초한 출원을 말하는데, PCT에서 규정한 서류를 특허청에 제출하면 특허출원을 원하는 국가를 단순히 지정하기만 하면, 지정된 국가에서도 특허출원이 인정된다. 그러나 PCT를 출원했다고 세계적으로 효력을 발생시키는 국제특허를 받을 수 있는 것은 아니며, 특허를 받기 위해서는 소정 기간 내에 각 국가에 진입해야 한다.

이 과정에서 각 국가는 해당 기술에 대한 선행기술 조사를 실시하여야 하고 그때 각 국가의 심사진행 정보를 확인하기 위해 WIPO(World International Patent Operation)의 패밀리정보를 참고해 패밀리특허 정보 조사를 통해 해외에서의 심사 자료 및 등록현황을 참고한다(이처영, 2001).

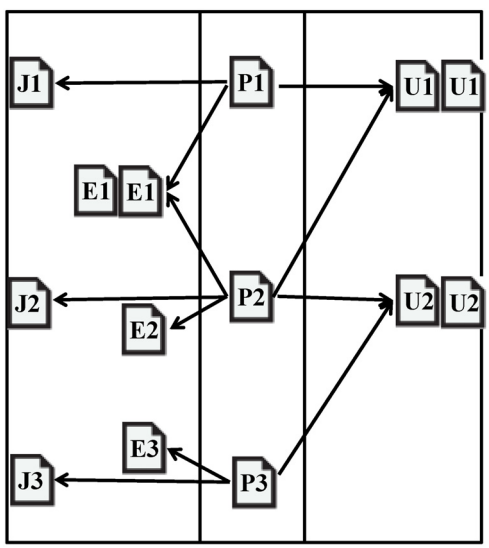
선행기술조사는 국제특허공개공보 검색을 기초로 진행하는데 WIPO에서 정보를 제공하고 있고 국내 심사정보 이력도 WIPO에 근거하여 제공한다.

2.1.6 패밀리특허

하나의 특허출원은 여러 개의 우선권 주장을 할 수 있으며, 또한 하나의 특허출원은 하나의 우선권 주장을 기초 출원으로 여러 개의 우선권 주장 출원으로 확장될 수 있다. 이런 특질로부터 패밀리 특허는 발생된다.

<그림 1>에서 볼 수 있는 바와 같이 우선권 주장의 최우선 출원인 P1, P2, P3가 있을 경우, 이 출원의 집합들에서 파악되는 패밀리특허의 개수는 일반적으로 3개로 파악된다. P1은 U1, J1, E1들과 함께 하나의 패밀리를 형성하고, P2는 U1, U2, J2, E1, E2들과 하나의 패밀리를 형성한다.

마지막으로 P3는 U2, E3, J3들과 패밀리를 구성하고 있다. 그러나 OECD는 이러한 일반적인 개념보다도 패밀리의 개념을 더욱 넓히고 있다. P2와 P1은 E1과 U1을 매개체로 하여 직간접적으로 연결되어 있다.



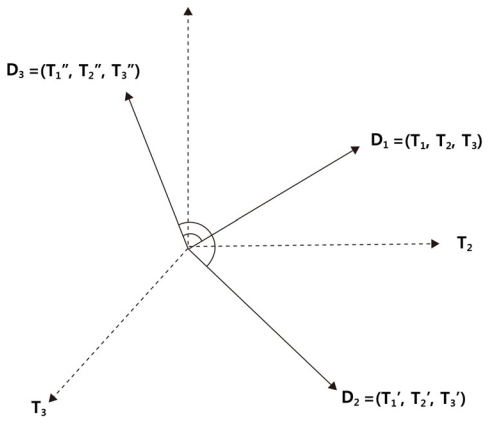
〈그림 1〉 패밀리특허 형성 예시

이 경우까지도 하나의 패밀리로 간주한다는 것이다(한국특허정보원, 2008). 따라서 하나의 특허출원 문서의 패밀리특허 수가 2개에서 많게는 수천 개의 특허문서 그룹이 될 수도 있다. 본 연구에서는 이들 간에 유사성을 비교해 우선 참고 할 수 있는 추천 시스템을 제안한다.

문서 간 유사성을 측정하는 방법을 제안했던 선행연구들을 살펴보면, 김우주 등(2016)은 Word2vec을 활용하여 동의어를 학습하고 유사한 문서를 찾는 방법을 제안하였으며, 심준식, 김형중(2017)은 LDA 토픽 모델링을 활용하여 판례에 숨어있는 동의어를 찾아 판례 검색에 도움을 주는 방법을 제안하였다.

2.2 Vector Space Model

출원문서의 유사성을 비교하기 위해선 문서의 비정형 텍스트데이터를 전처리 하고 벡터화 하는 절차가 필요한데 이를 위해 벡터공간모델을 사용한다. 즉 벡터란 방향성과 같은 길이(크기)를 가진 객체를 말하는데, 정보검색에서 벡터공간모델은 개별 용어와 그 용어들의 집합이라 볼 수 있는 검색내용과 문서색인을 벡터공간에 표현 할 수 있다는 가설을 기반으로 한다. 벡터공간 모델은 자연어처리(NLP)에서 주로 사용되는 방법으로, 같은 컨텍스트(Context)에 있는 단어는 같은 의미(Semantic meaning)를 공유한다고 가정한다. 이런 가정은 언어학에서 분포가설(Distributional Hypothesis)이라 하며, 이런 가정을 통해 벡터 공간 모델에서 문서(Document)는 벡터로 임베딩(embedding) 되어 표현되고, 개별차원(dimension)은 각각 단어(term)에 대응된다. <그림 2>에 표시한 벡터공간에서 두 벡터간 Cosine 각도를 이용해 유사도를 계산한다(Salton *et al.*, 1975).



〈그림 2〉 Ideal Document Space
(Salton *et al.*, 1975)

2.3 Word Embedding

단어 임베딩(Word Embedding) 이란 텍스트를 구성하는 하나의 단어를 수치화 하는 방법의 일종으로 어떻게 하면 컴퓨터가 텍스트를 이해할 수 있을까? 라는 물음에서 시작되었다. 언어를 수치적인 방식으로 표현하는 One-hot encoding으로는 단어와 단어 간의 관계를 알 수 없고 0, 1로 모든 단어를 표현하기에는 너무 많은 차원의 벡터가 생성되는 문제를 해결하고자 발전된 word Embedding은 설정한 k개의 차원으로 대상을 대응 시켜 표현하며 이 대응을 임베딩(Embedding)이라 한다. 하나의 정보가 여러 차원에 분산되어 표현되며 하나의 차원은 여러 속성들에 버무려진 정보를 내포한다. 이는 앞서 설명한 분포가설로 one hot encoding의 단점을 개선하기 위한 방법으로 적은 차원으로 대상을 일반화 하는 능력으로 발전해 왔다.

2.4 Perceptrons

One hot encoding 의 단점을 개선한 NNLM은 word embedding의 방법으로 뇌를 모사한 신경망 알고리즘(Perceptrons)을 언어학에 발전시킨 것으로, 신경 세포의 입력을 받고 출력을 내보내는 함수와 같은 형태를 적용시킨 방법이다. Perceptrons의 특징은 입력값이 들어왔을 때 입력값을 바로 출력하지 않고 가중치(Weight)를 곱하여 내보낸다. 이 과정을 앞먹임(feed-forward)이라고 하고 이렇게 가

중치를 곱한 입력 값의 결과를 출력하는데, 이를 실제 목표 출력치와 비교하여 다음 입력 때는 출력치가 목표치에 근접 할 수 있도록 가중치를 조절하는 과정을 거친다. 이 과정을 역전파(back-propagate) 라고 하며 이 앞먹임과 역전파 과정을 많은 데이터를 입력하면서 번갈아가며 진행하다 보면 가중치가 데이터에 일반적(general)으로 적용되게 되는데, 이를 학습(learning)이라 한다. 하지만 perceptrons의 학습은 훈련 데이터가 선형 분리 문제여야 하고 작은 학습률의 단점이 있다. 단층 신경망의 단점을 보안하기 위해 다층 신경망이 발전하였고 이 과정을 언어에 적용하여 발전한 것이 NNLM이다. 언어 모형(language model)이란, 단어가 문서에서 출현하는 과정을 확률 프로세스로 보고, 특정 위치에 특정 단어가 출연할 확률이 얼마나 되는지를 계산하기 위한 것이다. 이 후, 이 NNLM 방법이 RNNLM이라는 방법으로 발전하여 현재 Word embedding의 대표적 방법인 Word2vec으로 발전하게 되었다.

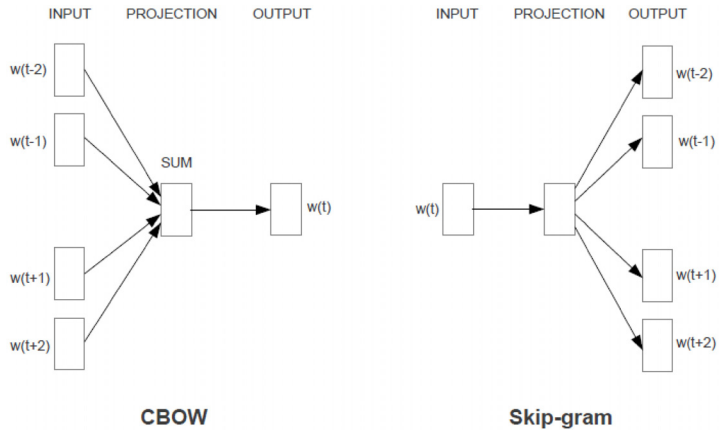
2.5 Word2vec

Word2vec은 단어의 위치와 의미를 내포하는 벡터의 형태로 Word embedding 하는 가장 대표적인 모델로서 앞서 설명한 벡터공간모델과 동일한 가정을 기반으로 한다. 즉, Word2vec은 단층 신경망을 이용한 신경망 언어모델(Neural language model)로 각각의 단어를 벡터로 표현하는데 <그림 3>와 같이 그 과정에서 특정 단어 주변에 오는 단어들의 집합을 이용해 특정 단어를 추측하는 CBOW 모델과 특정 단어 주변에 올 수 있는 단어를 유추하는 Skip-gram 모델을 이용해, 각각의 문맥에서 유추할 수 있는 w(단어)의 각각의 확률을 최대화하는 방법으로 학습한다. 따라서 유사한 단어들은 유사한 벡터의 위치를 가지게 되고 유사도가 높아지게 된다. 각각의 단어 벡터공간의 위치는 각 단어의 관계를 나타내고 있기 때문에 단어의 상관관계는 벡터의 거리로 표현 할 수 있다(Mikolov *et al.*, 2013).

본 연구에서 사용하는 gensim 패키지의 Word2vec은 Skip-gram과 CBOW의 알고리즘을 선택할 수 있고 제Ⅲ장 모델생성에서 그 파라미터 선택 실험을 진행한다.

2.6 LDA(Latent Dirichlet Allocation)

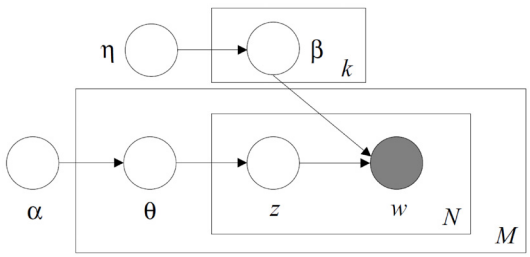
텍스트 마이닝 분석에서 토픽 모델링으로 가장 많이 활용되고 있는 LDA(Latent Dirichlet Allocation)는 기존의 LSA(Latent Semantic Analysis)(Deerwester *et al.*, 1990)와 PLSA(Probabilistic Latent Semantic Analysis)(Hofmann, 2013) 등의 약점을 보안한 방법으로 문서 내에 잠재되어 있는 주제(Topic)들을 추



〈그림 3〉 CBOW algorithm and Skip-gram algorithm of Word2vec

론 하는 생성확률모델(Generative probabilistic model)이다. LDA는 <그림 4>와 같이 특정 문서가 문서 내 여러 주제 중 각 주제에 속할 확률분포와 특정 단어가 각 주제에 속할 확률분포를 깁스 샘플링(Gibbs Sampling)을 활용해 구하고, 새로운 문서에 포함된 단어를 통해 해당 문서의 주제를 추론하는 모델이다. 이미 관찰된 변수(observed variable)를 통해 각각의 확률을 계산하여 토픽을 생성하는 사후 추론방법이다.

즉 각각의 요소를 가진 단어 벡터는 집합을 이루는 하나의 문서벡터로 표현될 수 있고 벡터들간의 코사인(Cosin) 각으로 유사도를 계산한다.



〈그림 4〉 LDA algorithm(Blei *et al.*, 2003)

2.7 선행연구와의 차별점

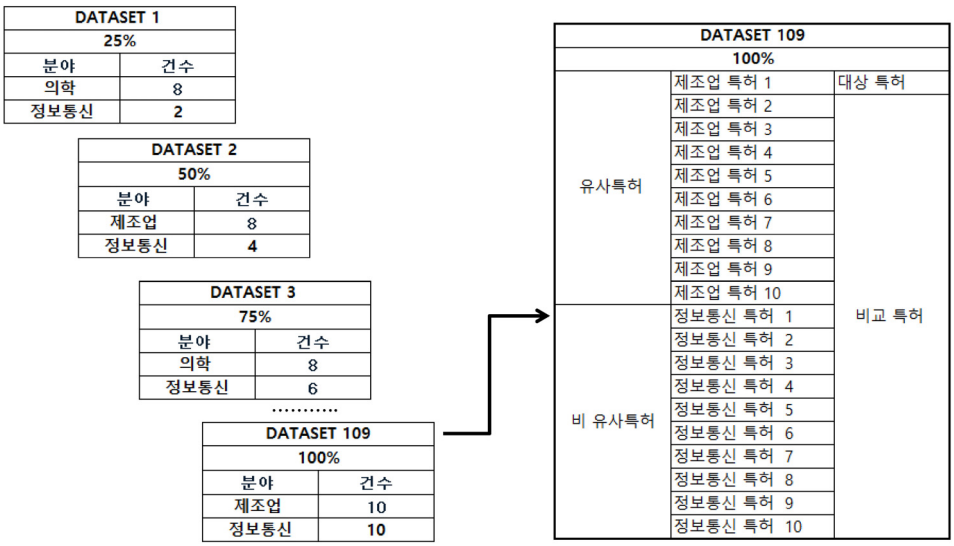
본 연구는 일반적인 말뭉치로부터 word embedding 작업을 수행한 모델을 사용한 선행연구들과는 달리, 선행연구들에서 많이 다루어지지 않은 해외특허문서를 대상으로 하였다는 점, 그리고 해외특허문서가 갖고 있는 시대적, 민족적, 언어적, 문화적 상황 및 각 분야에 따른 동음이의어, 이음동의어 특징을 반영해 각 도메인에 적합한 word embedding 값을 학습하는 모델을 구현하여 유사 특허문서를 탐색하였다는 점에서 선행연구들과의 차별점을 갖는다. 또한 특허문서는 명칭, 초록과 같이 함축적인 표현을 하는 부분과, 청구항과 같은 세부사항을 상세하게 기술하는 부분의 유형별로 각각의 단어가 다른 의미와 강도를 가지기 때문에 한 부분만 사용하거나, 전체를 통합하여 사용할 경우 특허의 특질을 제대로 반영하지 못할 수 있다는 선행연구들의 한계점을 보완하였다는 점에서 차별점을 갖는다.

Ⅲ. 모델 생성

3.1 데이터 수집

본 연구에서 분석을 위해 사용한 데이터는 국내 특허청에서 제공하는 국제특허심사정보 사이트의 패밀리 특허출원문서이다. 이는 유럽특허청 DOCDB를 근거로 하고 있어 DOCDB 자료를 검색할 수 있는 유럽특허청 EPO(European Patent Office)에서 제공하는 패밀리 특허출원문서를 대상으로 하였다. 언어는 국제특허에서 표준으로 하는 영어를 사용하였으며, 모델의 유사특허 문서 분류 성능 테스트를 위해 비 유사 문서로 식별 가능하도록 검색 대상을 정보통신 분야, 의학 분야, 제조업 분야의 특허출원 문서 382건의 명칭, 초록, 청구항 부분을 수집하였다. 본 연구는 본 연구에서 제안한 방법이 유사한 문서들 속에 비 유사 문서들이 포함되어 있을 경우 이 비 유사 문서들을 얼마나 잘 식별하는지 살펴보기 위해 다음과 같이 실험을 진행하였다. 우선, 정보통신, 의학, 제조업 분야에서 수집한 각 특허문서의 패밀리특허문서 수는 1~50개로 다양하게 이루어져 있었는데, 비 유사 문서의 판별력 평가를 위해 특정분야의 특허문서를 다른 분야의 특허문서와 25%, 50%, 75%, 100%의 비율로 섞어서 dataset을 구성한 결과 <그림 5>와 같이 109개의 dataset이 생성되었다.

각각의 dataset에서 1번째 특허와 유사하지 않은 문서를 분류하는 실험을 109개의 dataset에 대해 분류의 적합성을 재현율(recall)과 정밀도(precision)라는 척도로 측정하였으며, 매 학습 시 word embedding 값이 <그림 6>과 같이 변할 수 있고 이것이 결과에 영향을 미칠 수 있기 때문에 결과의 신뢰도를 높이고자 동일한 조건에서 총 10회씩 반복하여 그 평균값이 높은 방법을 채택하였다.



〈그림 5〉 DATASET 구성

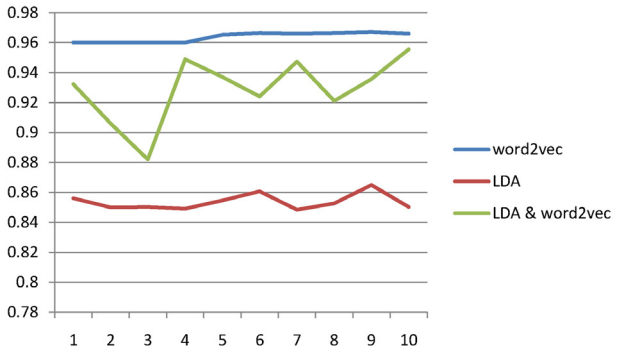
3.2 데이터 전처리

수집된 텍스트 데이터를 수치형 자료로 표현하기 위해 가장 기본적인 단위인 '토큰(token)'으로 분리하는 '토큰화(tokenization)' 작업을 수행하였으며, 문서집단에서 고빈도 출현 단어들이 존재하고 대부분의 텍스트 마이닝 기법들은 이러한 단어의 출현 빈도에 기반을 둔다. 이것은 지프의 법칙(Zipf's law)이 적용되는데 지프의 법칙에 따르면 어떠한 자연어 말뭉치 사용 빈도는 해당 단어의 순위에 반비례한다. 또한 가장 사용 빈도가 높은 단어는 다음 단어보다 약 두 배 빈도가 높으며, 그 다음 단어보다는 빈도가 세 배 높으며 지프에 법칙에 따르면 미국 표준 영어 말뭉치의 경우, 가장 빈도가 높은 단어는 정관사 'the'이며 문서의 7%의 빈도를 차지한다고 한다. 특허 문서의 Claims 같은 경우는 반복적으로 Claims이라는 단어를 표기하게 되고 이에 단어 영향을 줄 수 있다(Han *et al.*, 2011). 특허 문서의 분석 결과 노이즈를 최소화하기 위해 불필요한 조사, 공통적으로 등장하는 명사, 기호들을 모아 불용어 집합을 만들어 제거하였다. 마지막으로 형태소 분석(Stemming)을 통해 어형이 변형된 단어로부터 접사를 제거하고 그 단어의 어간을 분리하는 작업을 수행하였다.

3.3 모델 선택

3.3.1 1단계: 분류 실험

제Ⅱ장 선행연구에서 설명하였던 LDA, Word2vec 알고리즘과 LDA와 Word2vec을 통합하여 함께 사용한 방법을 각각 109개의 dataset에 10회씩 반복하여 분류실험을 진행하였다. Word2vec은 n차원의 vector에 각 단어를 표현하는 학습을 진행하여야 하는데 패밀리 특허문서는 동일한 우선권을 주장하는 그룹이기 때문에 해당 분야에서 유사하게 사용하는 단어와 표현을 학습할 수 있는 데이터가 보장이 되며 LDA 또한 gibbs sampling을 통해 n차원의 vector에 각 단어와 문서에 대한 주제 분포를 학습할 수 있다. 본 논문에서는 실험에 대한 적합성(Relevance)의 기준으로 관련 있는 모든 문서를 잘 찾아내는가를 의미하는 재현율(Recall)과 모델이 찾은 문서 중 관련된 문서가 얼마나 있는가를 의미하는 정밀도(Precision), 그리고 두 지표를 조화 평균하여 종합적으로 반영한 F-1 Score를 사용하였다.



〈그림 6〉 1단계 분류 실험의 F-1의 그래프

word2vec										
구분	1회	2회	3회	4회	5회	6회	7회	8회	9회	10회
Precision	97%	97%	97%	97%	97%	98%	98%	98%	98%	98%
Recall	95%	95%	95%	95%	95%	95%	95%	95%	95%	95%
F-1 Score	96%	96%	96%	96%	97%	97%	97%	97%	97%	97%

LDA										
구분	1회	2회	3회	4회	5회	6회	7회	8회	9회	10회
Precision	98%	96%	97%	95%	97%	97%	95%	97%	98%	96%
Recall	76%	76%	76%	77%	76%	77%	76%	76%	78%	76%
F-1 Score	86%	85%	85%	85%	85%	86%	85%	85%	86%	85%

LDA & word2vec										
구분	1회	2회	3회	4회	5회	6회	7회	8회	9회	10회
Precision	97%	94%	91%	97%	97%	95%	96%	96%	97%	97%
Recall	90%	87%	86%	93%	91%	90%	93%	88%	90%	94%
F-1 Score	93%	91%	88%	95%	94%	92%	95%	92%	94%	96%

〈그림 7〉 1단계 분류 실험 결과

<그림 6>, <그림 7>은 1단계의 결과를 보여준다.

1단계의 실험결과를 살펴보면, Word2vec을 사용한 모델이 10회에 걸친 실험에서 모두 95% 이상의 높은 Precision, Recall, F-1 Score 값을 보인 반면, LDA를 사용한 모델은 Recall 값이 77% 수준으로 상대적으로 낮게 나타났으며, LDA & Word2vec를 복합적으로 사용한 모델은 Precision, Recall, F1-Score 값이 93% 수준으로 비교적 높게 나타났으나, Word2vec의 분류 정확도가 더 높고, 일관된 결과를 보여주어 본 논문은 Word2vec을 유사 특허를 찾기 위한 알고리즘으로 선택하였다.

3.3.2 2단계: 최적 parameter 선택

실험을 통해 선택된 Word2vec은 제Ⅱ장에서 설명하였듯이 여러 알고리즘을 포함하고 있는데 실험도구로 사용한 gensim.word2vec은 여러 Parameter를 선택할 수 있고 그 선택에 따라 결과가 달라질 수 있기 때문에 최적의 Parameter선택 과정이 필요하다. 대표적인 Parameter 값은 <표 1>과 같다.

〈표 1〉 대표적인 Parameter 설명

size	단어 벡터의 차원
window	문장 내에서 현재 단어와 예측 단어 사이의 최대 거리
min_count	최소 단어 빈도수
workers	사용 스레드 개수
sg	0: CBOW , 1: Skip-gram
alpha	초기 학습속도
seed	난수 생성기의 시드
iter	코퍼스에 대한 반복 횟수(epochs)

window = 2										
구분	1회	2회	3회	4회	5회	6회	7회	8회	9회	10회
Precision	87%	87%	87%	87%	87%	87%	87%	87%	87%	87%
Recall	99%	99%	99%	99%	99%	99%	99%	99%	99%	99%
F-1 Score	92%	92%	92%	92%	92%	92%	92%	92%	92%	92%

window = 8										
구분	1회	2회	3회	4회	5회	6회	7회	8회	9회	10회
Precision	78%	79%	79%	78%	78%	78%	78%	78%	78%	78%
Recall	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
F-1 Score	88%	88%	88%	88%	88%	88%	88%	88%	88%	88%

sg = 0 (CBOW)										
구분	1회	2회	3회	4회	5회	6회	7회	8회	9회	10회
Precision	73%	73%	73%	73%	73%	73%	73%	73%	73%	73%
Recall	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
F-1 Score	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%

sg = 1 (Skip-gram)										
구분	1회	2회	3회	4회	5회	6회	7회	8회	9회	10회
Precision	91%	91%	91%	91%	91%	91%	91%	91%	91%	91%
Recall	99%	99%	99%	99%	99%	99%	99%	99%	99%	99%
F-1 Score	95%	95%	95%	95%	95%	95%	95%	95%	95%	95%

〈그림 8〉 최적 Parameter 선택 실험 결과

1단계에서 사용한 dataset을 이용하여 동일한 조건에서 실험한 결과 대표적으로 영향을 많이 주는 Parameter는 <그림 8>과 같다.

큰 차이를 보인 Window는 학습 시 주변단어의 개수를 몇 개까지 고려할지에 대한 Parameter로 2~10개로 변경하여 실험한 결과 Window 사이즈가 커지면 Recall 수치는 높아지나 F-1 Score는 Window size가 2일 때 더 좋은 성능을 보여 Window size는 2로 선택하였고, Word2vec의 알고리즘 중 Skip-gram이 평균 95%의 높은 성능을 보여 Skip-gram을 다음 단계의 실험에서 사용하였다.

3.3.3 3단계: 학습 데이터 선택

특허문서의 각 부분은 다른 특성을 가지고 있고 그에 따른 중요도가 다를 수 있기 때문에 명칭, 초록, 청구항의 비율을 항목별 10%씩 변화시켜 각각 0%~100%로 적용하여 <표 2>와 같이 66가지 경우의 수에 대해 모두 실험을 진행하였다.

실험한 결과는 작은 차이를 보였지만 명칭(Invention) 20%, 초록(Abstract) 30%, 청구항(Claims) 50% 비율이 가장 좋은 결과를 보이는 것으로 나타났다.

〈표 2〉 3단계 학습 데이터 실험 비율 예시

명칭	초록	청구항
100%	0%	0%
90%	10%	0%
90%	0%	10%
80%	20%	0%
80%	0%	20%

Ⅳ. 모델 적용

생성한 모델을 적용하기 위해, 광범위한 특허분야 중 ICT(Information and Communications Technologies) 정보통신 분야의 패밀리건수가 10~100인 100건의 출원으로 선정하여 100건의 특허문서 dataset을 구성하여 10회의 실험을 진행하였다.

명칭, 초록, 청구항을 구분하지 않는 2단계 적용 모델과 3단계 적용 모델의 성능을 비교한 결과는 <그림 9>와 같다.

모델을 생성할 시에는 전혀 다른 분야의 문서를 대상으로 하였고 모델을 적용할 시에는 같은 분야의 문서를 비교하였기 때문에 모델 생성 시 측정한 성능 보다는 전체적으로 낮은 결과를 보였지만

2단계 적용 모델										
구분	1회	2회	3회	4회	5회	6회	7회	8회	9회	10회
Precision	97%	96%	97%	96%	96%	97%	97%	96%	97%	96%
Recall	92%	92%	92%	92%	92%	92%	92%	91%	92%	92%
F-1 Score	94%	94%	94%	94%	94%	95%	94%	94%	94%	94%

3단계 적용 모델										
구분	1회	2회	3회	4회	5회	6회	7회	8회	9회	10회
Precision	81%	81%	81%	81%	81%	81%	81%	81%	81%	81%
Recall	95%	95%	95%	95%	95%	95%	95%	95%	95%	95%
F-1 Score	88%	88%	88%	88%	88%	88%	88%	88%	88%	88%

〈그림 9〉 적용 모델 실험 결과

같은 분야의 특허문서를 비교 하였을 경우에도 전체적으로 90% 정도의 높은 성능을 보였다.

특히, 본 연구의 모델은 심사를 판단하는 것이 아니라 유사도가 높은 문서를 추천 해주는 모델이기 때문에 모든 문서를 찾아내는 재현율이 중요한데, 본 연구에서 제안하는 3단계 모델은 재현율이 95%로 2단계 모델의 92%보다 높은 결과를 보였다.

Ⅴ.결 론

본 연구는 특허의 출원 및 심사 활동에 도움을 주고자 검색 특허의 패밀리특허문서를 수집 및 전처리한 후 Word2vec 알고리즘을 활용하여 가장 유사한 특허문서를 추천하는 모델을 제안하였다. 제안한 모델을 실제로 구현하였으며, 실험을 통해 모델의 성능을 평가하였다. 실험결과 약 95%의 높은 분류 정확도를 보이는 것으로 나타났다.

본 연구는 동일한 우선권을 주장하는 해외패밀리특허문서를 대상으로 해당 분야에서 유사하게 사용하는 단어와 표현을 학습하였고 Word2vec과 LDA의 각 장점을 극대화하려는 실험을 통해 더 좋은 결과를 얻었다는 점, 그리고 모든 문서 비교를 같은 방법론으로 적용하는 것이 아니라 각 특성에 맞게 적용하고 명칭, 초록, 청구항의 중요도에 따라 비율을 다르게 적용하여 모델의 성능을 더욱 높일 수 있었다는 점에서 학술적 시사점이 있다.

또한, 본 연구는 유사하다고 판단된 특허문서에 대한 심사 이력을 신속하게 제공하여, 패밀리 특허조사를 필요로 하는 심사관의 업무 부담감을 줄여 심사의 품질을 향상시키고, 출원인으로 하여금 효율적인 특허검색이 가능하게 도와준다는 점에서 실무적 시사점이 있다.

본 연구의 한계점은 데이터 수집의 어려움으로 특허문서의 명칭, 초록, 청구항 이외의 다른 항목들에 대한 실험을 추가하지 못하였다는 점이며, 이는 향후 연구에서 더욱 다양한 정보를 활용함으로써 보완하고자 하며, 이를 통해 모델의 성능을 더욱 개선시키고자 한다.

참 고 문 헌

[1] 강지호, 김종찬, 이준혁, 박상성, 장동식, “계층적 인용관계분석을 통한 선행기술 탐색방법론”, *한국지능시스템학회논문지*, 제27권, 제1호, 2017, pp. 72-78.

[2] 고광수, 정원교, 신영근, 박상성, 장동식, “텍스트 마이닝을 이용한 특허정보검색 개발에 관한 연구”, *한국산학기술학회논문지*, 제12권, 제8호, 2011, pp. 3677-3688.

[3] 김갑조, 윤다혜, 황종환, 선동주, “특허 토픽 모델링과 성장주기곡선을 통한 유망기술 발굴”, *한국지능시스템학회논문지*, 제27권, 제4호, 2017, pp. 357-363.

[4] 김우주, 김동희, 장희원, “Word2vec을 활용한 문서의 의미 확장 검색방법”, *한국콘텐츠학회논문지*, 제16권, 제10호, 2016, pp. 687-692.

[5] 심준식, 김형중, “LDA 토픽 모델링을 활용한 판례 검색 및 분류 방법”, *전자공학회논문지*,

제54권, 제9호, 2017, pp. 67-75.

[6] 이광희, 고순주, 김방룡, 전황수, 박광만, 석왕헌, 홍재표, *ICT 유망 기술의 생태계 및 산업 경쟁력 분석*, 한국전자통신연구원, 2015.

[7] 이원상, 손소영, "빅데이터 기술을 활용한 대용량 삼극특허 분석 기반의 기술융복합 패턴 예측", *대한산업공학회추계학술대회논문집*, 2013, pp. 1153-1170.

[8] 이처영, *생명공학 특허 전략*, 대광서림, 2001.

[9] 특허청, *특허청 2018년 주요 정책추진 방향*, 2018.

[10] 한국특허정보원, *Patent21*, 통권제80호, 한국특허정보원, 2008.

[11] Blei, D. M., A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation", *Journal of Machine Learning Research*, Vol.3, 2003, pp. 993-1022.

[12] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R., "Indexing by latent semantic analysis", *Journal of the American Society for Information Science*, Vol.41, No.6, 1990, pp. 391-407.

[13] Han, J., J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*, Elsevier, 2011.

[14] Hearst, M. A., "Untangling text data mining", In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, Association for Computational Linguistics, 1999, pp. 3-10.

[15] Hofmann, T., "Probabilistic latent semantic analysis", *ArXiv preprint arXiv:1301.6705*, 2013

[16] Mikolov, T., K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space", *arXiv preprint arXiv*, 2013, pp. 1301-3781.

[17] Salton, G., A. Wong, and C. S. Yang, "A vector space model for automatic indexing", *Communications of the ACM*, Vol.18 No.11, 1975, pp. 613-620.

Information Systems Review
Volume 22 Number 1
February 2020

LDA Topic Modeling and Recommendation of Similar Patent Document Using Word2vec

Apgil Lee* · Keunho Choi** · Gunwoo Kim***

Abstract

With the start of the fourth industrial revolution era, technologies of various fields are merged and new types of technologies and products are being developed. In addition, the importance of the registration of intellectual property rights and patent registration to gain market dominance of them is increasing in oversea as well as in domestic. Accordingly, the number of patents to be processed per examiner is increasing every year, so time and cost for prior art research are increasing. Therefore, a number of researches have been carried out to reduce examination time and cost for patent-pending technology. This paper proposes a method to calculate the degree of similarity among patent documents of the same priority claim when a plurality of patent rights priority claims are filed and to provide them to the examiner and the patent applicant. To this end, we preprocessed the data of the existing irregular patent documents, used Word2vec to obtain similarity between patent documents, and then proposed recommendation model that recommends a similar patent document in descending order of score. This makes it possible to promptly refer to the examination history of patent documents judged to be similar at the time of examination by the examiner, thereby reducing the burden of work and enabling efficient search in the applicant's prior art research. We expect it will contribute greatly.

Keywords: Patent Document, Similar Patent Document, LDA Topic Modeling, Text Mining, Word2vec

* Ph.D. Student, Department of Business, Hanbat National University
** Assistant Professor, Department of Business & Accounting Hanbat National University
*** Corresponding Author, Associate Professor, Department of Business & Accounting Hanbat National University

◐ 저 자 소 개 ◑

이 앞 길 (leeapgil@gmail.com)

현재 대전에 소재한 국립한밭대학교 일반대학원 경영학과 박사과정에 재학 중이다. 관심분야는 빅데이터 분석, 데이터마이닝, 딥러닝 등이다.

최 근 호 (keunho@hanbat.ac.kr)

현재 대전에 소재한 국립한밭대학교에서 경영회계학과 조교수로 재직하고 있다. 고려대학교 경영학과에서 박사 학위(MIS 전공)를 받았으며, 근로복지공단 근로복지연구원에서 데이터 분석 업무를 총괄하는 책임연구원으로 근무하였다. 주요 관심분야는 추천 시스템, 의료 빅데이터 분석, 딥러닝, 머신러닝, 데이터 마이닝 등이다.

김 건 우 (gkim@hanbat.ac.kr)

현재 대전에 소재한 국립한밭대학교에서 경영회계학과 부교수로 재직하고 있다. 연세대학교 공과대학에서 컴퓨터 사이언스를 전공하였으며 고려대학교 경영학과에서 석사를 졸업하고 동대학에서 박사 학위를 수여하였다. 현재 한국창업학회 부회장을 맡고 있으며 ICT플랫폼학회 빅데이터분과 위원장을 맡고 있다. 그 외 다수의 학회에서 편집위원 및 이사로서 활동하고 있다. 주요 관심분야는 비즈니스 온톨로지 모델, 빅데이터 분석 및 핀테크 기술 및 전략 등이다.

논문접수일 : 2019년 08월 01일 **게재확정일 : 2019년 09월 27일**
1차 수정일 : 2019년 09월 20일