공학석사 학위논문

특허 데이터를 통한 유망 기술 예측
: 유망 기술의 특징을 기반으로

Predicting Emerging Technologies through Patent Data:
Based on Characteristics of Emerging Technologies

2022년 2월

서울과학기술대학교 일반대학원
데이터사이언스학과

전 우 진

특허 데이터를 통한 유망 기술 예측
: 유망 기술의 특징을 기반으로

Predicting Emerging Technologies through Patent Data:
Based on Characteristics of Emerging Technologies

지도교수 금영정

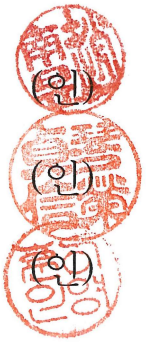
이 논문을 공학석사 학위논문으로 제출함
2022년 2월

서울과학기술대학교 일반대학원
데이터사이언스학과

전 우 진

전우진의 공학석사 학위논문을 인준함
2022년 2월

심사위원장 조남욱 (인)
심사위원 금영정 (인)
심사위원 이영훈 (인)



요 약

제 목 : 특허 데이터를 통한 유망 기술 예측: 유망 기술의 특징을 기반으로

미래사회 변화를 주도하는 기술 혁신을 파악하고, 유망 기술에 대한 지속적인 투자는 국가와 기업 차원에서 산업 경쟁력 향상을 위한 중요한 요소로 여겨진다. 따라서 유망 기술 예측은 산업과 기업 전반에서 큰 관심사로 간주되어 왔으며, 특히 특허 데이터를 사용하여 유망 기술을 예측하는 다양한 연구가 진행되어 왔다. 그러나 기존 연구는 주로 특허의 구조적 부분에 집중하여 유망 기술을 예측하였을 뿐 아니라, 유망 기술의 주요 특징을 고려하지 않고 연구를 진행하였다. 그러나 특허의 품질을 평가할 땐 중추적인 역할을 하는 정보는 특허의 비구조적 부분이기 때문에 해당 정보를 고려할 필요가 있으며, 동시에 유망 기술의 특징을 고려하여 기술 유망성의 다면적 특징도 고려할 필요가 있다. 이에 본 연구는 기술의 유망성에 대한 고찰을 바탕으로, 특허의 비구조적 부분과 구조적 부분을 통합하여 특허의 유망성을 예측하고자 한다. 본 연구는 특허의 구조적 부분을 활용한 기존 연구들을 바탕으로 특허 데이터로부터 서지학적 지표를 추출하고, 이에 더해 특허의 비구조적 부분인 텍스트 관련 지표를 제안하여 유망 기술의 5가지 특징에 따라 맵핑한 뒤, 특허의 유망성을 예측하였다. 특허 데이터는 PatentView API를 통해 수집한 14,992개의 헬스케어 기술 관련 특허를 사용하였다. 특허 데이터로부터 미리 정의한 설명변수를 추출 및 계산하였고, 머신러닝으로 특허의 유망성을 예측하였다. 실험 결과, 본 연구에서 제안한 특허 텍스트 지표 중 같은 기간에 특허의 초록에서 쓰이는 의미 있는 키워드의 빈도가 특허의 유망성을 결정하는 데 중요한 요인이라는 것을 확인할 수 있었다. 본 연구에서 제시하는 유망 특허 예측 프레임워크는 기업이 투자하기 적합한 특허 및 기술을 식별하는 데 도움이 될 것으로 기대된다.

목 차

표 목 차

그림목차

수식목차

Ⅰ. 서 론

1. 연구 배경

기술 혁신이 미래사회 변화의 중요한 동인으로 고려되면서 유망 기술에 대한 지속적인 투자는 국가와 기업 차원에서 산업 경쟁력 향상을 위한 중요한 요소로 여겨진다. 이때 한정된 투자 자원을 사용하여 연구개발의 효율을 높이기 위해서 유망 기술을 선제적으로 파악하여 해당 기술에 집중하는 전략이 필요하다(Kwon & Geum, 2020). 따라서 유망 기술에 대한 사전적 발굴은 투자 관련 전략을 수립하는 것을 목표로 하는 산업과 정부 관계자들에게 큰 관심사이다(Rotolo et al., 2015). 그러나 기술의 발전 속도가 매우 빠를 뿐 아니라 기술적 성공은 다양한 요인에 의해 결정되기 때문에 유망 기술을 사전에 파악하는 것은 매우 어렵다고 알려져 있다(Kyebambe et al., 2017).

유망기술의 발굴에 대한 기존 연구들은 대부분 특허 데이터를 활용하고 있다. 특허 데이터는 기술 혁신에 대한 막대한 정보를 제공하고(Kyebambe et al., 2017), 기술력을 측정하는 데에 널리 사용되어왔을 정도로 (Daim et al., 2006), 기술의 대용지표로써 이용될 수 있기 때문이다(Kim & Lee, 2015). 특히 다수의 기존 연구들은 특허 피인용 정보가 기술의 경제적 가치를 측정하는 데에 유용하다는 것을 증명한 바 있기 때문에 특허의 피인용 지표는 전반적인 기술가치를 측정하는 데 폭넓게 활용되어 왔다(Lerner et al., 1994; Narin et al., 1987). 하지만 이러한 기존 연구들은 사후적으로 특허의 과거 성과, 영향 또는 결과를 측정하는 형태의 평가로 제한되었기 때문에 그 결과가 미래지향적인 방향을 고려하지 않았다 (Lee et al., 2016)는 한계가 있다. 즉, 특허의 정보만을 바탕으로 미래의 특허 유망성을 사전적으로 살펴보기 위한 노력이 필요하다.

먼저 특허는 특허 출원인, 특허 인용, 특허 분류 등으로 구성된 구조적 부분과 초록, 청구항, 설명 등으로 구성된 비구조적 부분으로 나누어 볼 수 있다(Kim & Lee, 2017). 이전 연구들은 특허가 승인된 이후에 곧바로 사용될 수 있는 특허의 구조적 데이터들을 활용하여 이러한 단점을 극복하고자 하였다. Lee et al.(2018)은 특허의 구조적 데이터만을 활용하여 머신러닝과 딥러닝 모델을 통해 초기 단계에 있는 유망 기술을 식별하고자 하였다. Kyebambe et al.(2017)은 특허의 구조적 데이터를 활용하여 특허 문서를 벡터로 임베딩하여 특허들을 클러스터링하고, 코사인 유사도를 이용하여 유망 기술의 클러스터를 예측하였다. 하지만 두 연구 모두 특허의 구조적 데이터만을 이용하여 유망 기술을 식별 및 예측하고자 하였고, 특

허 그 자체의 내용을 담고 있는 비구조적 데이터를 활용하진 않았다는 한계가 존재한다. 특허의 비구조적 데이터를 구조적 데이터와 같이 활용하면 유망 특허를 예측할 때 성능의 개선을 가져올 수 있다(Lee et al., 2018; Niemann et al., 2017). 특허의 비구조적 데이터 중 하나인 텍스트 정보는 특허가 승인되고 바로 사용할 수 있으므로 시간에 따른 과대평가나 과소평가가 이루어지지 않는다는 장점이 있기 때문이다(Ju & Sohn, 2015). 이러한 단점을 극복하기 위해 특허의 비구조적 데이터를 기술 융합이나 유망 특허를 예측하는 데에 활용한 연구도 존재한다. Kim & Sohn(2020)은 IPC 쌍을 입력받아 구조적 데이터와 비구조적 데이터를 임베딩 벡터로 변환하여 머신러닝 모델을 통해 기술 융합을 예측하였다. Chung & Sohn(2020)의 연구에서도 특허의 구조적 데이터와 비구조적 데이터를 각각 임베딩 벡터로 변환하여 CNN과 Bi-LSTM을 기반으로 하는 딥러닝 모델을 학습하였고 특허의 가치를 예측하였다.

하지만 특허의 비구조적 데이터를 활용하여 유망 기술을 예측하고자 하는 이러한 기존 연구들은 두 가지 한계점이 존재한다. 첫 번째는 대부분의 기존 연구들이 입력변수로 활용가능한 특허 항목 자체에 초점을 맞추는 접근을 취하고 있기 때문에, 유망기술 자체가 가진 특성에 대한 폭넓은 고찰이 부족하다는 점이다(Lee et al., 2018). 유망기술은 일반적으로 ‘Novelty’, ‘Fast Growth’, ‘Coherence’, ‘Prominent impact’, ‘Uncertainty and ambiguity’ 등의 5가지 특성을 가지는 것으로 알려져 있다(Rotolo et al., 2015). 진정한 의미의 유망기술 예측은 유망기술이 가지는 특성을 명확히 이해하고, 이를 분석 과정에 충실히 반영하는 과정이 선행되어야 한다. 두 번째 한계는 특허의 비구조적 데이터를 활용하는 대부분의 연구가 해당 데이터를 벡터화해 이용하는 데 그쳤다는 점이다. 비구조적 데이터에서 텍스트를 단순 벡터 표현으로 임베딩하면 해당 특허가 가진 특징을 명확히 표현하지 못한다는 한계가 존재하고, 이에 특정 기술군에서 중요한 키워드를 놓칠 수 있는 단점이 존재한다. 이는 기술이 유망성을 가지기 위해 어떤 특성을 가져야 하는지에 대한 실질적인 시사점을 제공하지 못한다는 한계가 있다.

따라서 본 연구에서는 유망 기술 예측을 위해 특허의 구조적 데이터만 사용하는 과정을 개선하여 다양한 기술적 시사점을 얻을 수 있는 대용량 텍스트 정보인 비구조적 데이터를 활용하고자 한다(Gerken & Moehrle, 2012). 또한, 특허 데이터로 유망 기술의 특징에 기반을 둔 지표를 개발하고 설명변수로 활용하여 유망 기술의 다면적 특징을 모두 고려한 예측을 수행하고자 한다.

2. 연구 목적

이에 본 연구는 특허의 구조적 데이터와 비구조적 데이터를 유망 기술의 특징을 기반으로 머신러닝 모델에 적용한 유망 특허 예측 모델을 개발하고, 더 나아가 유망 기술을 예측하고자 한다. 이를 위해 유망기술의 특징을 바탕으로 특허의 유망성에 영향을 끼칠 만한 구조적 및 비구조적 평가기준을 도출하였다. 이를 바탕으로 다수의 머신러닝 모델을 통해 지도 학습을 수행하고 어떤 특허가 유망한 특허가 되는지 그 특징을 알아본다. 이 연구는 기존에 주로 구조적 데이터로 이뤄지던 특허 분석과 달리 다양한 비구조적 데이터를 활용함으로써 유망 기술 예측의 신뢰도와 정확성을 높일 수 있다.

3. 연구 구성

본 연구는 다음과 같이 구성된다. 2장에서는 본 연구의 배경이 되는 기존 연구를 살펴본다. 3장에서는 연구 방법을 제시하고, 연구 흐름에 관하여 서술한다. 그 다음 4장에서는 제시한 연구 방법의 사례 연구를 수행하여 얻은 결과를 살펴보고 관련 시사점을 도출한다. 마지막으로 5장에서는 본 연구의 결론을 서술한다.

Ⅱ. 문 헌 연 구

1. 이론적 배경

1) 유망 기술의 특성에 관련한 연구

유망 기술의 특징을 정의한 연구는 다양하다. Martin(1995)의 연구에선 유망 기술은 ‘Prominent impact’, ‘Scope and coverage’, ‘Development effort and developers’ capabilities’, ‘Science-intensity’의 특징을 가진다고 주장하였다. 해당 연구에선 ‘Prominent impact’는 유망 기술의 영향이 기술 분야를 넘어 사회, 경제 시스템에 영향을 주는 특성을 말한다. ‘Scope and coverage’는 유망 기술에 담겨있는 기술군의 범위를 뜻하며, 해당 연구에서 가장 중요한 특성으로 주장하였다. ‘Development effort and developers’ capabilities’는 유망 기술을 개발하는 과정에 들어간 노력과 개발자의 능력을 의미한다. ‘Science-intensity’는 기술이 얼마나 과학에 기반하고 있는지를 뜻한다. Day & Schoemaker(2000)은 유망 기술이 ‘Prominent impact’, ‘Scope and coverage’, ‘Uncertainty and ambiguity’, ‘Development effort and developers’ capabilities’, ‘Novelty’, ‘Science-intensity’, ‘Coherence’와 같은 특징을 보인다고 하였다. ‘Uncertainty and ambiguity’는 유망한 기술이 개발된 의도대로 좋은 결과만 가져오는 것이 아니라 기술의 오용 등 의도하지 않은 결과도 가져온다는 것을 의미한다. ‘Novelty’는 유망 기술의 참신함을 뜻한다. 그리고 ‘Coherence’는 유망 기술이 지속적으로 사용이 되는 것을 의미한다. 해당 연구는 이러한 특징들을 바탕으로 유망 기술을 새로운 산업을 창출하거나 기존 산업을 변화시킬 수 있는 잠재력을 가진 과학 기반 혁신으로 정의하였다. 또한, Porter et al.(2002)의 연구에선 ‘Prominent impact’, ‘Uncertainty and ambiguity’, ‘Development effort and developers’ capabilities’을 유망 기술의 특징으로 보았다. Cozzens et al.(2010)의 연구에선 ‘Prominent impact’, ‘Uncertainty and ambiguity’, ‘Science-intensity’, ‘Growth speed’을 유망 기술의 특징으로 정의하였는데, 이때 ‘Growth speed’는 기술의 성장 속도를 의미하며 유망 기술은 그 속도가 빠르다고 보았다. Small et al.(2014)의 연구에선 유망 기술이 ‘Novelty’, ‘Growth speed’의 두 가지 특성으로 정의된다고 하였

다. Rotolo et al.(2015)는 이러한 연구들을 종합하여 최종적으로 유망 기술의 특징을 ‘Novelty’, ‘Fast Growth’, ‘Coherence’, ‘Prominent impact’, ‘Uncertainty and ambiguity’ 5가지로 정리하였다.

Table 2.1 유망 기술의 특징에 대한 기존 연구

	Prominent impact	Scope and coverage	Uncertainty and ambiguity	Development effort and developers' capabilities	Novelty	Science-intensity	Coherence	Growth speed	Fast Growth
Martin (1995)	✓	✓		✓		✓			
Day & Schoemaker (2000)	✓	✓	✓	✓	✓	✓	✓		
Porter et al.(2002)	✓		✓	✓					
Cozzens et al.(2010)	✓		✓			✓		✓	
Small et al.(2014)					✓			✓	
Rotolo et al.(2015)	✓		✓		✓		✓		✓

Lee et al.(2018)은 이러한 유망 기술의 특징을 기반으로 특허의 서지학적 정보를 맵핑하였는데, 이때 *fast growth*와 *coherence*는 제외하였다. 그리고 해당 데이터를 머신러닝 모델로 학습하여 유망 기술을 식별하고 그 성능을 측정하였다. 하지만 해당 연구는 유망 기술의 특징 5가지를 모두 사용하지 않았고, 특허의 텍스트 정보를 활용하지 않았다.

2) 특허를 이용한 유망 기술 예측

특허는 기술에 대한 정보를 보유한 원천으로써 상업적 가치를 지니기 때문에 특허 분석은 유망 기술을 예측하는 데 매우 중요하다(Choi & Hwang, 2014). 이러한 이유로 특허를 통해 유망 기술을 예측하거나 특허의 품질을 분류한 다양한 연구들이 있는데, 이를 입력변수의 특성에 근거하여 다음과 같이 두 가지 유형으로 나눌 수 있다.

(1) 특허의 구조적 데이터를 활용한 유망 기술 예측

Breitzman & Thomas(2015)는 가까운 미래에 나타날 흥미로운 기술이나 유망 기술을 식별할 수 있는 유망 클러스터 모델을 제안하였다. 해당 연구는 기존에 비슷한 모델이 미국에서 출원된 특허에만 집중한 것을 개선하여 다양한 국가에서 출원된 특허를 모두 고려한다는 특징이 있다. 연구에서 제안한 유망 클러스터 모델은 먼저 인기 있는 특허를 식별한다. 이때 인기 있는 특허란 두 가지 조건을 만족해야 하는데, 먼저 대상 특허의 피인용 빈도 중 최근에 나온 특허의 빈도가 높아야 하고, 그 비율이 대상 특허의 피인용 빈도의 대부분을 차지해야 한다. 이러한 정의를 바탕으로 기준 연도에 유망 특허와 상호 인용의 관계를 맺고 있는 특허들을 유망 클러스터를 식별하였다.

Érdi et al.(2013)은 새로운 특허 인용 네트워크 내에 새로운 클러스터의 출현 여부를 이용하여 새로운 기술의 출현을 감지하는 모델을 개발하였다. 해당 연구에서 네트워크의 노드는 특허 벡터를 통하여 구성하였다. 이때 특허 벡터는 미리 선별된 36개의 기술군으로부터 대상 특허의 피인용 빈도의 합계로 계산하였다. 이렇게 구성된 네트워크를 기간별로 분석하여, 미국특허청(USPTO)보다 새로운 기술 영역의 출현을 빠르게 식별하고자 하였다. 하지만 해당 연구는 기존에 존재하는 기술 클러스터가 후속 연도의 네트워크에서 높은 비중을 가진다는 점과 피인용 빈도를 기반으로 하므로 충분한 인용수를 얻기 전까지 시간이 필요하다는 단점이 존재한다.

Kwon & Geum(2020)의 연구에서는 축적된 기술지식 품질을 고려하여 유망 기술을 식별하고자 하였다. 단순히 축적된 기술지식의 양뿐만 아니라 그 품질을 고려하고자 대상 특허가 인용하고 있는 인용 특허의 빈도도 고려하였다. 해당 연구는 종속변수로 대상 특허의 특정 기간의 피인용 빈도를 사용하였는데, 이는 시간에 따라 누적되는 피인용의 특성을 고려한 것이다. 그리고 설명변수로 인용 특허의 개수, 독립 청구항의 개수 등 총 17개의 변수를 사용하였다. 마지막으로 수집한 데이터로 머신러닝 모델을 통해 성능 측정을 수행하였다.

(2) 특허의 비구조적 데이터를 활용한 유망 기술 예측

하지만 특허의 구조적 데이터는 시간에 따라 달라지는 특징이 있기에 그 한계가 있다(Kim & Sohn, 2020). 실제로 특허의 품질을 평가할 땐 중추적인 역할을 하는 정보는 특허의 비구조적 부분이기 때문에 해당 데이터를 활용할 필요가 있다(Girthana & Swamynathan, 2018; Anne et al., 2018; 임소라 & 권용진, 2017).

Kim & Sohn(2020)은 기존에 쓰였던 특허의 구조적 데이터에 더해 특허의 비구조적 데이터를 활용하여 기술 융합을 예측하였다. 해당 연구에선 수집한 특허 쌍을 기간별로 나누고, 링크 예측, 서지학적 정보, 의미적 유사도를 통해 특허 IPC쌍의 융합 패턴을 나타낼 수 있는 융합 벡터를 생성하였다. 이때 특허의 비구조적 부분인 특허 제목, 초록, 청구항을 Doc2vec으로 임베딩하여 해당 IPC를 나타낼 수 있는 벡터로 변환하여 사용하였다. 최종적으로 얻은 세 가지 벡터를 합쳐 융합 벡터로 변환하고 이를 머신러닝 기법을 통하여 새로운 기술 융합을 예측하였다.

Chung & Sohn(2020)은 특허의 텍스트 정보를 딥러닝 모델의 설명변수로 사용하여 연구에서 미리 정의한 특허 등급으로 분류하는 프레임워크를 제시하였다. 먼저 수집한 특허의 등급을 피인용의 빈도를 바탕으로 A, B, C 세 등급으로 정의하였다. 그리고 특허의 텍스트 정보 중 초록과 청구항을 각각의 단어 임베딩 벡터로 변환하여 행렬 형태로 만들었다. 특허의 텍스트 정보만을 활용한 예측 결과와 비교를 위해 특허의 서지학적 정보도 예측에 추가로 사용하였다. 예측 모델은 CNN과 Bi-LSTM을 결합한 딥러닝 모델을 사용하였으며, 실험 결과에서 텍스트 정보만을 활용하여 특허의 등급을 예측한 경우가 가장 높은 성능을 보였다.

이러한 연구들과 동시에 텍스트 마이닝을 이용하여 특허의 기술 키워드를

기반으로 유망 기술을 예측한 연구들도 존재한다. Geum et al.(2013)은 특허-키워드 빈도 행렬을 사용하여 novelty 탐지 기법에 적용해 기존에 알려지지 않은 새롭고 유망한 특허를 식별하였다. Joung & Kim(2017)은 TF-IDF로 기술 키워드를 정의하였고, 정의된 키워드들의 연관성을 동의어, 상위어, 하위어 등의 관계로 구분한 뒤, 기간별 클러스터 분석을 통하여 유망 기술을 식별하였다.

2. 방법론적 배경

본 연구에서는 유망 특허를 예측하고, 특허의 유망성에 영향을 주는 중요한 변수를 확인하기 위하여 다양한 머신러닝 방법론을 활용하였다. 대표적으로 로지스틱 회귀, 의사결정나무, 랜덤 포레스트, LightGBM, XGBoost, AdaBoost, Gradient Boosting을 사용하였고, 여러 모델의 예측 결과를 통합하여 voting을 이용해 최종 예측을 하는 Voting 분류 모델도 추가하여 실험에 사용하였다. 머신러닝 모델들은 각자 장단점이 확실하므로 유망 특허 예측이라는 분류 문제에 있어 가장 높은 성능을 보이는 모델을 선택하고자 한다. 본 연구에서는 유망기술의 예측 자체도 중요하지만 유망기술에 영향을 주는 요인에 대한 고찰도 매우 중요하기 때문에 텍스트 자체를 직접 임베딩해서 사용하는 딥러닝 모델은 고려하지 않았다.

Ⅲ. 연 구 방 법

1. 연구 프레임워크

본 연구의 흐름은 Fig. 3.1에서 알 수 있듯이 먼저 데이터 수집, 특허 지표 정의 및 전처리, 특허의 유망성 측정, 실험 결과 검증 및 중요 변수 확인, 실험 결과 분석의 순서로 구성된다.

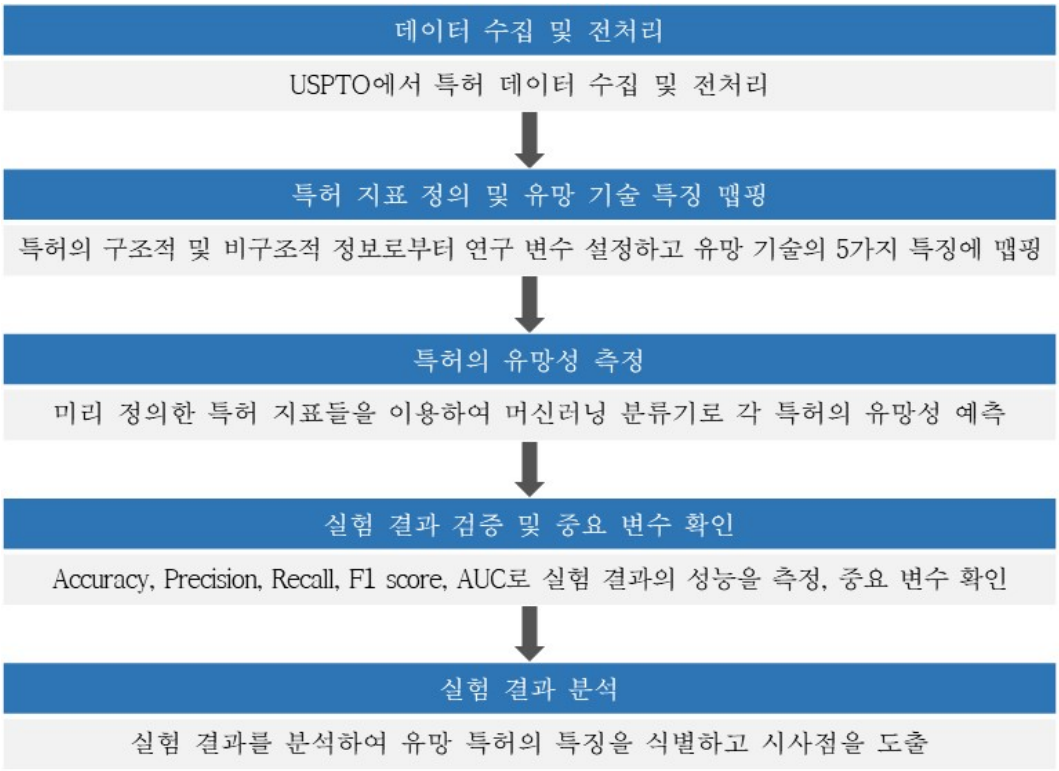


Fig. 3.1 연구 프레임워크

먼저 미국특허청(USPTO)에서 헬스케어 산업 분야에 대한 특허를 수집하고, 각 특허는 미리 설계한 지표 정의에 따라 데이터 전처리를 수행한다. 해당 단계가 완료된 특허 데이터에서 구조적 및 비구조적 정보로부터 미리 정의한 설명변수를 추출하고 필요한 전처리를 수행한다. 정의한 설명변수들은 해당 특성에 따라 유망 기술의 5가지 특징인 ‘radical novelty’, ‘fast growth’, ‘coherence’, ‘prominent impact’, ‘uncertainty and ambiguity’에 각각 맵핑한다. 이후 훈련 데이터로 머신러닝 모델을 학습하고 검증 데이터로 그 성능을 측정한다. 이때 성능은 Accuracy, Precision, Recall, F1-score 그리고 AUC로 평가한다. 마지막으로

실험 결과를 통해 유망 특허의 특징을 식별하고 관련 시사점을 도출한다.

2. 데이터 수집

특허의 수집은 다른 나라의 특허청에 비해 그 규모나 경제적 가치가 큰 미국특허청(USPTO)에서 수집한다. 따라서 두 가지 과정을 통해 미국특허청에서 특허를 수집한다. 먼저 Google Patent의 Advanced Search 기능을 이용해 특정 산업에 대한 키워드를 검색한다. 이때 검색 옵션에서 Patent Office는 US, Language는 English, Status는 Grant, Type은 Patent로 설정하여 검색을 진행한다. 검색된 특허의 고유번호를 이용하여 PatentView API로부터 각 특허에 대한 데이터를 수집한다. 설명변수에 사용할 특허의 고유번호, 제목, 출원일, 특허 종류, 초록, CPC(Cooperative Patent Classification), 출원자, 출원자의 특허 개수, 인용 특허 번호, 인용 특허 출원일, 피인용 특허 번호, 피인용 특허 출원일, 심사관 고유 번호, 심사관 직위를 수집한다. 이때 CPC란 특허 분류체계를 의미하는데, 기존 특허 분류체계인 IPC보다 하나 더 많은 section을 가지고 16개 더 많은 subclasses를 가지고 있기 때문에 해당 분류체계는 기술을 더욱 세분화하여 분류한다. 본 연구에서는 CPC의 section, classes, sub-classes, group까지만 사용한다(Fig. 3.2). CPC에 대한 전처리로 CPC sub-group을 제거하였다. 마지막으로 특허가 출원된 나라의 수인 Family Patent number는 PatentView API에서 지원하지 않으므로 Google Patent Advanced Search에서 각 특허의 고유 번호를 이용하여 해당 부분만 추가로 수집한다.

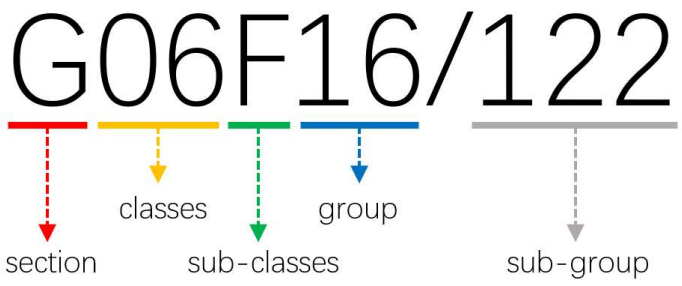


Fig. 3.2 CPC 예시

3. 유망 기술 특징 정의 및 변수 설정

1) 유망기술의 특징 정의

앞에서 서술하였듯이 유망 기술의 특성에는 ‘Scope and coverage’, ‘Development effort and developers’ capabilities’, ‘Science-intensity ’ 등의 다양한 특성이 존재한다. 본 연구에서는 이전에 정의되었던 유망 기술들의 특징을 정리하여 새롭게 정립한 Rotolo et al.(2015)의 5가지 유망기술 특성을 활용한다. 이 5가지 특징은 ‘radical novelty’, ‘fast growth’, ‘coherence’, ‘prominent impact’, ‘uncertainty and ambiguity’로 정의된다. 먼저 ‘radical novelty’란 유망 기술의 급진적인 참신함을 의미한다. 즉, 비슷한 목적을 달성하기 위해 이전에 사용되었던 기술과 비교하여 다른 기본 원리를 사용하여 주어진 기능을 수행하는 부분이라고 할 수 있다. ‘fast growth’란 유망 기술이 유망하지 않은 기술보다 성장 속도가 빠르다는 것을 의미한다. ‘coherence’란 기술의 지속성과 일관성을 나타내는 특징이다. 예를 들어, 한 기술 분야에서 여러 기술이 생기다가 시간이 지나며 특정 기술만 사용되는 경우가 그 기술의 지속성을 나타낸다고 할 수 있다. 또한, 한 기술 분야에서 다양한 용어가 사용되다가 용어 수가 감소하고 약어 또는 두문자어가 나타나는 경우도 기술의 지속성을 나타낸다. ‘prominent impact’는 유망 기술이 특정 산업 분야, 더 나아가서 사회경제 시스템에 광범위한 영향을 미치는 것을 뜻한다. 마지막으로 ‘uncertainty and ambiguity’는 유망 기술이 바람직한 결과만 가져오는 것이 아니라 의도하지 않은 결과도 가져올 수 있다는 불확실성을 의미한다.

2) 변수 설정

본 연구의 종속변수는 Forward Citation (FC)로, 대상 특허가 출원되고 5년 이내에 받은 피인용 빈도 합계로 정의된다(Petruzzelli et al., 2015). 특정 특허가 다른 특허들에서 인용되는 횟수를 나타내는 피인용은 특허의 시장가치와 밀접한 관련이 있고 특허의 가치를 평가하는데 널리 쓰이기 때문에 많은 연구에서도 특허의 유망성을 나타내는 지표로 쓰인다(Lerner et al., 1994; Haupt et al., 2007; Lanjouw & Schankerman, 2004). FC를 바탕으로 상위 10%의 특허를 유망 특허로, 나머지 특허를 유망하지 않은 특허로 라벨링을 진행한다.

설명변수는 유망기술의 특성인 ‘radical novelty’, ‘fast growth’, ‘coherence’, ‘prominent impact’, ‘uncertainty and ambiguity’을 포괄할 수

있도록 각 기준에 따라 다양하게 구성한다. 본 연구에서 구성한 설명변수는 다음과 같다.

(1) Radical novelty

Radical novelty는 유망 기술의 급진적 참신성, 즉 유망 기술이 기존 기술과 얼마나 다른가를 의미한다. 이를 설명하기 위한 특허지표는 다양한 관점에서 설명될 수 있으며, 본 연구에서는 radical novelty에 해당하는 특허지표를 prior knowledge, originality, radicalness, LOF score의 네 가지로 구성하였다. 먼저 Prior Knowledge는 대상 특허의 인용 빈도 합계다. 높은 수의 인용 특허를 가진 특허는 비교적 낮은 novelty와 상업적 가치를 가지는 경향이 있다고 알려져 있으므로 (Harhoff et al., 2003) 해당 지표를 통하여 기술의 novelty를 유추할 수 있다. 다음으로 Originality는 수식 (1)과 같이 대상 특허를 피인용하는 특허들의 cpc 비율을 제곱합 하여 1에서 뺀 값이다. 피인용 특허 중 cpc의 비율이 다양할수록 해당 변수의 값은 커질 것이고, 다양한 기술군에 영향을 끼쳤다고 볼 수 있다(Squicciarini et al., 2013).

$$1-\sum_{j\in S_B} CPC_j^2 \quad (1)$$

$S_B =$ Set of group level cpcs of forward citations

$CPC_j =$ Ratio of the cpc j of forward citations belongs to S_B

그리고 Radicalness는 수식 (2)와 같이 대상 특허의 인용 특허의 cpc 클래스 중 target patent의 cpc 클래스와 다른 것들의 비율이다. 특정 특허의 인용 특허에 다른 분야의 기술이 많이 포함되면 그 특허는 급진적이라고 할 수 있다(Squicciarini et al., 2013).

$$\sum_{i\in P} BCC_i/n_P \quad (2)$$

$P =$ Set of backward citation

$BCC_i =$ Number of backward citation i' s cpc which is not in target patent' s cpc

$n_P =$ Number of P

마지막으로 LOF score는 각 특허의 초록을 통해 계산된 LOF 점수다. LOF는 이상치 탐지 알고리즘 중 하나로써 이상치의 극단 정도를 수치값으로 표현하여 더욱 객관적인 해석이 가능한 알고리즘이다(Schubert et al., 2014). 주로 LOF 값이 1 이상일 때, 이상치라고 판단하며 값이 커질수록 정상 패턴과 차이가 크다고 본다(Kim & Lee, 2017). 따라서 특정 특허의 초록의 LOF score가 높을수록 그 특허가 novel할 가능성이 있다고 유추할 수 있다.

(2) Fast Growth

Fast Growth는 유망 기술이 유망하지 않은 기술보다 성장 속도가 빠르다는 것을 의미한다. 본 연구에서는 fast growth에 해당하는 특허지표를 Technology Cycle Time으로 구성하였다. Technology Cycle Time은 대상 특허의 인용 특허들의 중위 연령이다. Technology Cycle time이 짧을수록 기술군의 변화가 빠르다는 것을 의미하고 이는 특허 출원일로부터 지난 시간을 통해 측정할 수 있다(Bierly & Chakrabarti, 1996; Kayal & Waters, 1999).

(3) Coherence

Coherence는 유망 기술이 특정 산업 분야에서 지속성과 일관성을 나타낸다는 특징이다. 본 연구에서는 coherence에 해당하는 특허지표를 Irreplaceable Keyword in Claim, Irreplaceable Keyword in Abstract의 두 가지 텍스트 지표로 구성하였다. 먼저 Irreplaceable Keyword in Claim은 같은 기간 특허들의 청구항에서 TF-IDF 값 기준 상위 30개의 단어가 대상 특허의 청구항에서 나타나는 빈도를 구한 값이다. 이때 TF-IDF값은 수식 (3)과 같이 계산된다. 청구항과 같은 특허 텍스트 데이터는 기술 내용을 포함하고 특허 평가에서 중추적인 역할을 한다(Girthana & Swamynathan, 2018; Anne et al., 2018; 임소라 & 권용진, 2017). 따라서 청구항에서 특정 키워드의 평균 빈도가 높을수록 특정 기술군에서 사용하는 키워드가 보편화됨을 알 수 있다. 이와 비슷하게 Irreplaceable Keyword in Abstract는 같은 기간 특허들의 초록에서 TF-IDF 값 기준 상위 10개 단어의 단어가 대상 특허의 초록에서 나타나는 빈도를 구한 값이다. 일반적으로 특허의 청구항이 특허의 초록보다 길이가 기므로 각각 상위 기준값을 다르게 설정한다(Chung & Sohn, 2020).

$$w_{x,y} = tf_{x,y} \times \log(\frac{N}{df_x}) \quad (3)$$

$tf_{x,y}$ = Frequency of word x in patent text y

df_x = Number of patent text documents containing x

N = Total number of patent text documents

(4) Prominent Impact

Prominent Impact는 유망 기술이 특정 산업 분야, 더 나아가서 사회경제 시스템에 광범위한 영향을 미치는 것을 의미한다. 본 연구에서는 prominent impact에 해당하는 특허지표를 external knowledge, broad impact, examiner impact의 세 가지로 구성하였다. 먼저 External Knowledge는 대상 특허의 피인용 중 논문이나 책 같은 Non-Patent Literature의 개수이다. 특정 특허에서 사용된 Non-patent literature 인용수와 관련 기술의 가치는 밀접한 관계가 있으므로(Callaert et al., 2006) 특허의 가치를 유추하기에 적합하다. 다음으로 Broad Impact는 수식 (4)와 같이 대상 특허의 피인용의 cpc 클래스 중 target patent의 cpc 클래스와 다른 것들의 비율이다. Radicalness와 비슷하게 target 특허의 피인용의 cpc에서 target 특허의 cpc와 다른 것들이 많으면, target 특허가 광범위한 산업에 영향을 미친다고 볼 수 있으므로 'prominent impact'를 측정하기에 적합하다.

$$\sum_{i \in P} FCC_i / n_P \quad (4)$$

P = Set of forward citation

BCC_i = Number of forward citation i's cpc which is not in target patent's cpc

n_P = Number of P

마지막으로 Examiner Impact는 대상 특허의 피인용 중 특허 심사관이 인용한 빈도이다. Target 특허를 심사관이 forward cite 하는 횟수가 높을수록 그 특허가 비슷한 도메인에서 미치는 영향력이 높다고 볼 수 있다.

(5) Uncertainty and Ambiguity

Uncertainty and Ambiguity는 기술의 불확실성을 의미한다. 이는 성공에 대한 불확

실성 뿐 아니라 사회적 의미의 불확실성, 즉 유망 기술이 언제나 바람직한 결과만 가져오는 것이 아니라, 기술의 악용과 같은 의도하지 않은 결과도 가져올 수 있다는 불확실성 또한 포함하고 있다. 본 연구에서는 불확실성을 측정하기 위해 다양한 특허지표를 활용한다. 일반적으로 기업 및 개별 기술의 기술성과 관련되어 있는 것으로 알려진 여러 가지 지표들을 본 연구에서는 불확실성을 해소할 수 있는 지표로 정의하고 heterogeneity, under examination claim index, family patent, assignee patent, forward citations of assignee patent, examiner type의 여섯 가지로 구성하였다. 먼저 Heterogeneity는 출원인의 특허 중 대상 특허의 기술 분야(Health care)와 다른 특허의 비율이다. 출원인이 가진 특허 중 대상 특허와 다른 기술 분야의 특허 비율이 높을수록 해당 특허의 불확실성이 높다고 볼 수 있다. 다음으로 Under Examination Claim Index는 수식 (4)와 같이 대상 특허의 cpc 세트와 90% 이상 일치하는 특허의 청구항 평균 개수이다. 특허의 심사 과정에서 청구항이 삭제될 수가 있으므로, 특정 cpc군의 청구항 평균 개수가 많을수록 관련 cpc군에 아직 심사 중인 특허가 많다고 할 수 있고(Lanjouw & Schankerman, 2001), 그만큼 불확실성이 높다고 볼 수 있다.

$$\frac{\sum_{i \in p} c_i}{n_p} \quad (5)$$

$p =$ 대상 특허와 cpc가 90% 이상 일치하는 특허

$c_i =$ 특허 i의 청구항 개수

$n_p =$ number of p

셋째로 Family Patent는 대상 특허가 출원된 국가의 수다. 특허법은 국가별로 다르고(OuYang et al., 2011), 중요한 특허일수록 다양한 국가의 특허 사무소에 등록되기 때문에(Lee et al., 2018), 해당 지표가 작을수록 특정 특허의 불확실성은 크다고 볼 수 있다. 그다음 Assignee Patent는 출원인이 가진 특허 수이고, Forward citations of Assignee Patent는 출원인이 받은 피인용의 합계다. Assignee Patent와 Forward citations of Assignee Patent 모두 특허 출원인의 역량을 반영하는 지표다(Ernst, 2003). 두 지표의 값이 낮을수록 출원인의 기술적 역량이 평가되지 않은 상태이고, 출원인이 새롭게 개발한 기술의 불확실성이 높다는 가정을 할 수 있다. 마지막으로 Examiner Type은 특허의 심사에 assistant 심사관의 참여 여부를 통해 1 또는 0의 값으로 정해진다. 특허 심사 시, 경력이 5년 이상인 primary 심사관이

특허 승인에 대한 권한을 가지는데, 이때 assistant 심사관이 특허 문서를 검토하는 경우가 있다. 불확실성이 높은 특허일수록 assistant 심사관이 참여하는 경우일 가능성이 클 것이라는 가정에서 해당 변수를 'uncertainty and ambiguity'에 맵핑한다.

Table 3.1 특허 변수 목록

Category	유망 기술 특징	변수종류	변수명	변수설명	참조
Output	technology potential	Patent Indicator	FC	Forward citation over five years $FC\begin{cases}1, if\ FC is\ 90\%\ percentile\\0, otherwise\end{cases}$	Chung & Sohn(2020)
Input	radical novelty	Patent Indicator	Prior Knowledge (PK)	Number of backward citations	Lee et al.(2018)
		Patent Indicator	Originality (OG)	1 - (HHI of forward citations) (Herfindahl index on cpcs of forward citations)	Squicciarini et al.(2013)
		Patent Indicator	Radicalness (RD)	backward citation의 cpc 클래스 중 target patent의 cpc 클래스와 다른 것들의 비율	Squicciarini et al.(2013)
		Text Indicator	LOF score (LOF)	LOF score of each patents' abstract	
	fast growth	Patent Indicator	Technology Cycle Time (TCT)	Median age of backward citations	Bierly & Chakrabarti(1996), Kayal & Waters(1999)
	coherence	Text Indicator	Irreplaceable Keyword	claim에서 나타나는 특정 keyword(ex. Abbreviation)의 빈도 (특정 keyword: 같은 기간 특허들의	

			in Claim (IKC)	claim에서 tf-idf 값 기준 상위 30개 단어)	
		Text Indicator	Irreplaceable Keyword in Abstract (IKA)	abstract에서 나타나는 특정 keyword(ex. Abbreviation)의 빈도 (특정 keyword: 같은 기간 특허들의 abstract에서 tf-idf 값 기준 상위 10개 단어)	
	prominent impact	Patent Indicator	External Knowledge (EK)	Number of non-patent literature references	Callaert et al.(2006)
		Patent Indicator	Broad Impact (BI)	forward citation의 cpc 클래스 중 target patent의 cpc 클래스와 다른 것들의 비율	
		Patent Indicator	Examiner Impact (EI)	Number of forward citations made by examiner	
	uncertainty & ambiguity	Patent Indicator	Heterogeneity (HT)	Assignee의 특허 중 target patent와 다른 기술 분야의 특허 비율	Lee et al.(2018)
		Patent Indicator	Under Examination Claim Index (UECI)	Target patent와 같은 cpc군의 claim의 평균 개수	
		Patent Indicator	Family Patent (FP)	Number of Family Patents	
		Patent Indicator	Assignee Patent (AP)	Number of Assignees' Patents	
		Patent Indicator	Forward citations of Assignee Patent (FAP)	Number of forward citations of Assignees' Patents	
		Patent Indicator	Examiner Type (ET)	$ET\begin{cases}1, examined with assistant\\0, otherwise\end{cases}$	

4. 특허의 유망성 측정

변수 설정이 완료되면 머신러닝 모델을 통해 특허의 유망성을 예측한다. 종속변수가 대상 특허가 출원되고 5년 이내에 받은 피인용 합계이기 때문에 2017년부터 출원된 특허는 제외한다. 분류 모델에 입력하기 전 설명변수는 정규화를 통한 전처리를 수행한다. 이후, 데이터 세트를 8:2의 비율로 학습 데이터와 검증 데이터로 분리한다. 학습 데이터의 설명변수와 종속변수로 머신러닝 모델(로지스틱 회귀, 의사결정나무, 랜덤 포레스트, LightGBM, XGBoost, AdaBoost, Gradient Boosting, Voting 분류 모델)을 학습하고, 검증 데이터의 종속변수와 검증 데이터의 예측값을 비교한다. 이때 설명변수 중 Text indicator를 제외한 Patent indicator로만 다시 한 번 학습을 진행하고 모든 변수를 사용한 학습결과와 비교한다.

5. 실험 결과 검증 및 중요 변수 확인

검증 데이터에 대한 모델의 정확도를 평가하기 위해 본 연구에서는 수식 (6), (7), (8), (9)와 같이 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1-score, AUC(Area Under the ROC curve)를 사용한다. TP_i는 클래스 i에 대한 TP, TN_i는 클래스 i에 대한 TN, FP_i는 클래스 i에 대한 FP, FN_i는 클래스 i에 대한 FN이다 (Davis & Goadrich, 2006). 정확도(Accuracy)는 수식 (6)과 같이 계산되며 이는 분류 모델이 검증 데이터 중에서 올바르게 분류한 데이터의 비율을 뜻한다. 정밀도(Precision)는 수식 (7)과 같이 계산되며 이는 분류 모델이 Positive로 분류한 데이터 중에서 실제 데이터가 Positive인 경우의 비율을 뜻한다. 재현율(Recall)은 실제 Positive 데이터 중 모델이 Positive로 분류한 경우의 비율을 뜻하며 수식 (8)과 같이 계산된다. F1-score는 정밀도(Precision)와 재현율(Recall)의 조화 평균으로, 수식 (9)와 같다. ROC 커브는 TP와 FP의 분포가 겹치는 정도에 따라 결정되며 겹치지 않을수록 1에 가까워지고, 겹칠수록 0에 가까워진다. 해당 값이 1에 가까울수록 모델이 좋은 성능을 내는 것으로 해석할 수 있다.

Table 3.2 혼동 행렬

		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

$$Accuracy = \frac{TP_i + TN_i}{TP_i + FP_i + FN_i + TN_i} \quad (6)$$

$$Precision = \frac{TP_i}{TP_i + FP_i} \quad (7)$$

$$Recall = \frac{TP_i}{TP_i + FN_i} \quad (8)$$

$$F1 = 2 \times (\frac{\mathrm{Pr}ecision_i \times Recall_i}{\mathrm{Pr}ecision_i + Recall_i}) \quad (9)$$

Ⅳ. 사례 연구

본 연구는 프레임워크를 실제로 적용해보고 성능을 검증하기 위해 사례 연구를 수행하였다. 본 연구에서는 사례 연구의 대상 기술 분야로 헬스케어 기술 관련 특허 데이터를 활용하여 유망 특허를 식별하고자 한다.

한국바이오협회의 보고서에 따르면 글로벌 디지털 헬스케어 산업 규모는 2020년 약 1520억 달러였으며, 2027년까지 약 5080억 달러 규모로 성장할 것으로 전망하였다. 또 헬스케어 산업에 대한 투자도 확대되고 있다. 디지털헬스 기업에 투자된 규모는 2019년에 77억 달러에서 2020년에는 146억 달러로 약 2배 증가하였고, 2021년에는 상반기에만 이미 147억 달러가 투자되어 전년도의 투자 규모를 넘어섰다. 이처럼 글로벌 헬스케어 산업의 규모는 점점 성장하고 있으며, 이미 디지털 헬스케어 기업에 대한 투자 규모의 증가 추세도 이어지고 있다. 특히 소비자들을 대상으로 하는 헬스케어 서비스는 단순 건강관리 및 보조 수단을 넘어 질병 모니터링 등으로 고도화되는 경향을 보이기 때문에 관련된 유망 기술의 파악을 필수적이다.

이처럼 헬스케어 산업은 기술이 핵심 역할을 하는 대표적인 산업이다. 따라서 관련 기술의 개발을 위해 기업 간 경쟁은 더욱 치열해질 것이고, 시장 점유를 위한 유망 기술의 식별은 필수적인 부분이 되었다. 이에 본 연구는 미국특허청(USPTO)의 헬스케어 산업 분야의 특허를 활용하여 유망 특허를 분류한다. 또한, 그 성능을 측정하고 특허의 어떤 특징이 유망성에 영향을 주는지 분석하고자 한다.

1. 데이터 수집 및 전처리

헬스케어 산업에 대한 특허 데이터를 수집하기 위해 Google Patents의 Advanced Search 기능으로 ‘health care’ 키워드를 검색하였고, 2002년 1월부터 2021년 7월까지 출원된 헬스케어 관련 특허 번호 32,768개를 수집하였다. 수집한 특허 번호를 이용하여 PatentsView API로부터 관련 특허 데이터(32.768개)를 수집하였다. 해당 연구에서 제안하는 종속변수가 대상 특허가 출원되고 5년 이내에 받은 피인용 합계이기 때문에 2017년부터 출원된 특허는 제외하여 최종적으로 14,992개의 특허 데이터를 사용하였다. 정의한 설명변수의 계산을 위하여 특허 데이터에 대한 피인용 특허 68,590개, 인용 특허 358,699개를 추가로 수집하였다. 본 연구에서 제안한 변수 중 특허의 텍스트 관련 지표인 LOF, IKC, IKA를 포함한 데

이터 세트와 이를 포함하지 않은 13개의 서지학적 지표만을 사용한 데이터 세트를 생성한다. 최종적으로 생성된 설명변수의 기초적인 정보는 Table 4.1, Table 4.2와 같다. 이때 FC 값을 기준으로 유망 특허 라벨링을 진행하는데, FC의 상위 10% 값인 12를 기준으로 FC가 12 이상이면 유망 특허, 12 미만이면 유망하지 않은 특허로 라벨링을 하였다.

Table 4.1 특허 지표 기초 통계 (Numeric)

Indicator	Mean	Std	Min	0.25	0.50	0.75	Max
FC	4.96	13.86	0.00	0.00	1.00	3.00	420.00
PK	28.14	55.21	0.00	3.00	8.00	26.00	461.00
OG	0.81	0.21	0.00	0.72	0.86	1.00	1.00
RD	0.53	0.31	0.00	0.31	0.57	0.78	1.00
LOF	1.04	0.04	0.98	1.01	1.03	1.05	1.87
TCT	6667.77	2867.69	0.00	5341	7196.00	8400	16688
IKC	79.55	133.85	0.00	0.00	0.00	125	2307
IKA	5.93	5.15	0.00	2.00	5.00	8.00	44
EK	0.31	1.06	0.00	0.00	0.00	0.00	30.00
BI	0.44	0.37	0.00	0.00	0.44	0.78	1.00
EI	0.03	0.19	0.00	0.00	0.00	0.00	3
HT	0.75	0.33	0.00	0.61	0.93	0.98	0.99
UECI	7.33	7.62	0.00	0.00	8.00	9.94	135
FP	3.45	4.19	0.00	1.00	2.00	4.00	72.00
AP	6571.76	19662.94	0.00	11.00	271.00	2220.00	144597.00
FAP	2762.90	12083.17	0.00	5.00	62.00	628.00	68794.00

Table 4.2 특허 지표 기초 통계 (Binary)

Indicator	0	1
ET	9302 (62.0%)	5690 (38.0%)

2. 특허의 유망성 측정

1) 데이터 준비

본 연구의 프레임워크에 따라서 추출된 데이터 세트는 변수의 단위가 다르다는 문제가 존재한다. 이를 위한 전처리로 최소-최대 정규화를 데이터 세트에 대해 MinMaxScaler로 수행하였다. 이후 데이터 세트를 8:2의 비율로 학습 데이터와 검증 데이터로 분리하였다. 그리고 학습 데이터에 클래스의 불균형이 존재하기 때문

에 오버샘플링 기법을 통하여 클래스별 특허의 수를 맞춰주었다(Table 4.3).

Table 4.3 학습 및 검증 데이터 세트

특허의 유망성	학습 데이터	검증 데이터	Total
0	10,724개	2,682개	13,406개
1	1,269개 (오버샘플링 후 10,724개)	317개	1,586개
Total	11,993개	2,999개	14,992개

2) 머신러닝 모델 설정

머신러닝에 사용한 분류기 중 로지스틱 회귀, 의사결정나무, 랜덤 포레스트, LightGBM, XGBoost, AdaBoost, Gradient Boosting의 파라미터 설정은 각각 Table 4.4, Table 4.5, Table 4.6, Table 4.7, Table 4.8, Table 4.9, Table 4.10과 같다.

Table 4.4 로지스틱 회귀 모델 설정

설정	값
C	6866
Penalty	L2(Ridge)

Table 4.5 의사결정나무 모델 설정

설정	값
Criterion	entropy
max_depth	24
max_features	log2
min_samples_leaf	1
min_samples_split	2

Table 4.6 랜덤 포레스트 모델 설정

설정	값
max_depth	12
max_features	log2
min_samples_leaf	2
min_samples_split	3
n_estimators	300

Table 4.7 LightGBM 모델 설정

설정	값
boosting_type	dart
learning_rate	0.05
max_depth	15
min_child_samples	5
num_iterations	2000
num_leaves	16

Table 4.8 XGBoost 모델 설정

설정	값
base_score	0.5
booster	gbtree
learning_rate	0.1
max_depth	6
num_leaves	6

Table 4.9 AdaBoost 모델 설정

설정	값
base_estimator	DecisionTreeClassifier(max_depth=5)
learning_rate	0.1
n_estimators	500

Table 4.10 Gradient Boosting 모델 설정

설정	값
criterion	friedman_mse
learning_rate	0.1
loss	deviance
max_depth	15
max_features	log2
n_estimators	300
subsample	0.75

3) 머신러닝 학습 결과

특허의 텍스트 관련 지표가 특허의 유망성을 예측하는 성능에 영향을 미치는지 알아보기 위하여 Table 4.11과 같이 A, B 두 경우로 구분하여 사례 연구를 진행하였다. 모형 A는 특허의 서지학적 지표에 더해 본 연구에서 제안한 특허 텍스트 관련 지표를 추가하여 총 16개의 설명변수로 실험을 진행한 경우다. 모형 B는 모형 A에서 특허의 텍스트 관련 지표만 제외한 나머지 13개의 변수를 이용하여 실험을 진행한 경우이다. 각 모델의 학습결과는 Table 4.12, Table 4.13과 같다. 두 개의 모형을 비교한 결과, 10개의 모델 중 모형 A에선 Gradient Boosting 모델이 가장 좋은 성능을 보였고, 모형 B에선 XGBoost 모델이 가장 좋은 성능을 보였다. 모형 A에서 로지스틱 회귀 모델과 랜덤 포레스트 모델을 결합한 앙상블 모델이 유망 특허 클래스에 대한 재현율이 0.87로 가장 높았지만, 낮은 정확도와 정밀도를 고려하여 성능이 가장 좋은 모델이 되지 못하였다. 또한, 특허의 텍스트 관련 지표가 추가된 모형 A가 모든 지표에서 성능이 상승하였다.

Table 4.11 머신러닝 모델 학습 시나리오

유망 기술의 특징	A (구조적, 비구조적 변수 모두 활용)	B (구조적 변수만 활용)
Radical Novelty	Prior Knowledge (PK)	Prior Knowledge (PK)
	Originality (OG)	Originality (OG)
	Radicalness (RD)	Radicalness (RD)
	LOF score (LOF)	-
Fast Growth	Technology Cycle Time (TCT)	Technology Cycle Time (TCT)
Coherence	Irreplaceable Keyword in Claim (IKC)	-
	Irreplaceable Keyword in Abstract (IKA)	
Prominent Impact	External Knowledge (EK)	External Knowledge (EK)
	Broad Impact (BI)	Broad Impact (BI)
	Examiner Impact (EI)	Examiner Impact (EI)
Uncertainty & Ambiguity	Heterogeneity (HT)	Heterogeneity (HT)
	Under Examination Claim Index (UECI)	Under Examination Claim Index (UECI)
	Family Patent (FP)	Family Patent (FP)
	Assignee Patent (AP)	Assignee Patent (AP)
	Forward citations of Assignee Patent (FAP)	Forward citations of Assignee Patent (FAP)
	Examiner Type (ET)	Examiner Type (ET)

Table 4.12 모형 A 실험 결과 (구조적, 비구조적 변수 모두 활용)

Model	클래스	Accuracy	Precision	Recall	F1	AUC
LR	비유망	0.650	0.95	0.64	0.77	0.675
	유망		0.19	0.71	0.30	
DT	비유망	0.811	0.93	0.85	0.89	0.654
	유망		0.27	0.45	0.34	
RF	비유망	0.801	0.95	0.82	0.88	0.722
	유망		0.29	0.62	0.40	
LGBM	비유망	0.882	0.93	0.93	0.93	0.691
	유망		0.44	0.45	0.45	
XGB	비유망	0.898	0.93	0.96	0.94	0.666
	유망		0.52	0.37	0.44	
AB	비유망	0.891	0.92	0.96	0.94	0.632
	유망		0.47	0.30	0.37	
GB	비유망	**0.899**	0.92	**0.97**	**0.94**	0.633
	유망		**0.54**	0.30	0.38	
Ensemble (LR+RF+XGB)	비유망	0.812	0.96	0.82	0.89	**0.770**
	유망		0.32	0.72	**0.45**	
Ensemble (LR+GB)	비유망	0.880	0.94	0.93	0.93	0.692
	유망		0.44	0.45	0.44	
Ensemble (LR+RF)	비유망	0.691	**0.98**	0.67	0.79	0.769
	유망		0.24	**0.87**	0.37	

Table 4.13 모형 B 실험 결과 (구조적 변수만 활용)

Model	클래스	Accuracy	Precision	Recall	F1	AUC
LR	비유망	0.617	0.94	0.61	0.74	0.628
	유망		0.16	**0.64**	0.26	
DT	비유망	0.761	0.92	0.80	0.86	0.616
	유망		0.20	0.43	0.28	
RF	비유망	0.770	**0.95**	0.79	0.86	**0.705**
	유망		0.26	0.62	0.36	
LGBM	비유망	0.805	0.94	0.84	0.89	0.680
	유망		0.28	0.52	0.36	
XGB	비유망	**0.866**	0.92	**0.93**	**0.93**	0.644
	유망		**0.37**	0.36	0.36	
AB	비유망	0.859	0.92	0.93	0.92	0.608
	유망		0.32	0.29	0.30	
GB	비유망	0.851	0.92	0.91	0.92	0.633
	유망		0.32	0.36	0.34	
Ensemble (LR+RF+XGB)	비유망	0.832	0.93	0.87	0.90	0.672
	유망		0.31	0.47	**0.37**	
Ensemble (LR+GB)	비유망	0.831	0.93	0.88	0.90	0.652
	유망		0.29	0.43	0.35	
Ensemble (LR+RF)	비유망	0.752	0.94	0.77	0.85	0.693
	유망		0.24	0.62	0.35	

Table 4.14 모델별 변수 중요도

RF			XGB			LGBM		
입력변수	중요도	기술 특성	입력변수	중요도	기술 특성	입력변수	중요도	기술 특성
BI	21.71	prominent impact	BI	32.87	prominent impact	FP	3700	uncertainty & ambiguity
OG	19.51	radical novelty	EK	28.04	prominent impact	AP	2835	uncertainty & ambiguity
FAP	15.99	uncertainty & ambiguity	FP	25.42	uncertainty & ambiguity	FAP	2822	uncertainty & ambiguity
PK	8.22	radical novelty	FAP	20.52	uncertainty & ambiguity	PK	2798	radical novelty
AP	6.87	uncertainty & ambiguity	**IKA**	18.07	coherence	**IKA**	2781	coherence
TCT	5.45	fast growth	AP	14.71	uncertainty & ambiguity	BI	2678	prominent impact
HT	4.16	uncertainty & ambiguity	OG	13.50	radical novelty	OG	2439	radical novelty
RD	3.96	radical novelty	PK	12.16	radical novelty	TCT	2221	fast growth
LOF	3.46	radical novelty	UECI	8.83	uncertainty & ambiguity	UECI	1597	uncertainty & ambiguity
IKA	3.30	coherence	**IKC**	7.70	coherence	HT	1370	uncertainty & ambiguity

특허 품질의 예측에 있어 중요한 지표를 파악하기 위하여 모형 A의 랜덤 포레스트, XGBoost, LightGBM 모델로 변수 중요도를 측정하였다. Table 4.14는 각 모델에서 변수 중요도 상위 10개의 변수를 나타낸다. 먼저 랜덤 포레스트에서 가장 중요한 변수는 broad impact(BI)로 측정되었다. Broad impact(BI)는 특정 특허의 피인용 특허가 다른 산업 분야에 속한 비율이므로 자신이 속한 기술군보다 다른 기술군에 더 많은 영향을 미친 특허가 유망기술을 대변한다고 생각할 수 있다. Originality(OG)는 두 번째로 높게 나왔다. 따라서 다양한 기술군에 영향을 끼치는 특허일수록 기술의 유망성에 큰 영향을 미치는 것을 확인하였다.

그리고 XGBoost에서 가장 중요한 변수도 마찬가지로 broad impact(BI)로 측정되었다. 또한, 두 번째로 중요한 변수는 External Knowledge(EK)로 측정되었다. 두 변수 모두 기술의 광범위한 영향을 미치는 정도에 대한 변수이므로, 사회경제 시스템에 관한 기술의 영향력이 기술의 유망성을 결정하는 데에 큰 역할을 한다고 이해할 수 있다. 이를 통해 특허뿐만 아니라 다양한 기술문서를 인용하는 정도가 기술의 유망성에 큰 영향력을 미치는 것을 알 수 있다.

다음으로 LightGBM에서 중요하게 측정된 변수는 family patent(FP)와 assignee patent(AP)이다. 두 변수 모두 유망 기술의 불확실성과 모호성을 나타낸다. 이를 바탕으로 기술의 불확실성이 해당 기술의 유망성에 큰 영향을 미치는 것으로 확인되었다.

추가로 본 연구에서 제안한 특허의 텍스트 지표를 살펴보면 다음과 같다. 먼저 텍스트 지표 중 하나인 IKA가 모든 모델에서 중요한 변수로 나타났고, 나머지 텍스트 지표인 LOF와 IKC도 각각의 모델에서 중요한 변수로 나타났다. 이를 통해 특허의 초록에서 유의미한 기술 키워드의 유무가 특허의 유망성을 결정하는 데 영향을 주는 것을 확인할 수 있었다. Table 4.15는 16개의 설명 변수에서 중요 변수로 측정된 Table 4.14의 변수들이 속한 유망 기술의 특징별 비중을 수식 (10)과 같이 계산한 것이다. 이를 통해 'Prominent Impact'가 특허의 유망성에 가장 적은 영향을 미치는 것을 확인할 수 있었다.

Table 4.15 유망 기술 특징별 중요 변수 비중

유망 기술의 특징	비중
Radical Novelty	66.66%
Fast Growth	66.66%
Coherence	66.66%
Prominent Impact	44.44%
Uncertainty & Ambiguity	66.66%

$$\frac{\sum_{i \in F} v_i}{n_c \times n_v} \quad (10)$$

$v_i =$ 유망기술의 특징 i에 속하는 설명변수 v

$n_c =$ 변수 중요도 측정에 사용된 모델 개수

$n_v =$ 설명변수 전체 개수

또한, 유망한 특허와 유망하지 않은 특허의 특징을 모형 A의 의사결정나무를 시각화하여 파악하였다. 먼저 유망한 특허의 규칙은 Table 4.16과 같이 2개로 정리할 수 있었다. 첫 번째, PK가 8.5 이상, FAP가 17.5 이상, assistant 심사관 없이 출원되고 UECI가 0.02 이상인 특허가 유망한 특허가 된다. 두 번째, PK 8.5가 이하, IKA가 1.5 이상, BI가 0.05 이상, FAP가 12.5 이상인 특허가 유망한 특허가 된다는 결과가 나왔다.

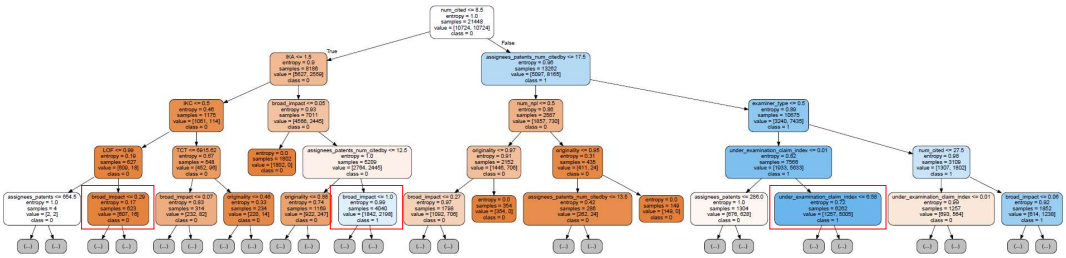


Fig 4.1 모형 A의 의사결정나무

반면 PK가 8.5 이하, IKA가 1.5 이하, IKC가 0.5 이하, LOF는 1 이상인 특허는 유망하지 않은 특허가 된다는 결과를 확인할 수 있었다. 모든 의사결정 규칙들의 lift 값이 1 이상이므로 유의미한 규칙들임을 알 수 있다.

Table 4.16 모형 A의 의사결정나무 규칙

구분	No.	Rule	Coverage	Lift
유망 특허	1	PK 8.5 이상, FAP 17.5 이상, assistatnt 심사관 없이 UECI가 0.02 이상인 특허	0.29	1.60
	2	PK 8.5 이하, IKA 1.5 이상, BI가 0.05 이상, FAP가 12.5 이상인 특허	0.19	1.09
비유망 특허	1	BC 8.5 이하, IKA 1.5 이하, IKC 0.5 이하, LOF가 1 이상인 특허	0.03	1.95

4) 실험 결과 분석

실험 결과, 특허의 서지학적 지표에 본 연구에서 제안한 3개의 특허 텍스트 관련 지표를 추가하여 예측한 결과가 서지학적 지표만 활용한 경우보다 더 나은 성능을 보였다. 이러한 결과는 특허 품질의 평가에서 특허의 비구조적 데이터가 중추적인 역할을 한다는 사실을 뒷받침한다(Girthana & Swamynathan, 2018; Anne et al., 2018; 임소라 & 권용진, 2017). 또한, 머신러닝 모델의 학습 과정에서 종속변수에 설명변수가 영향을 미치는 정도를 변수 중요도를 통해 측정하였다. 변수 중요도가 높은 순서대로 상위 10개의 변수들을 3개의 모델별로 추출하였는데, 본 연구에서 제안된 텍스트 관련 지표들이 공통적으로 종속변수 결정에 영향을 미치는 것을 알 수 있었다. 마지막으로 의사결정나무를 시각화 후 살펴본 결과, 유망 기술의 특징 중 ‘radical novelty’를 나타내는 지표 중 하나인 LOF가 높은 특허들이 유망하지 않은 특허가 된다는 결과가 나왔다. 일반적으로 기존 유망 기술의 특징을 토대로 생각해보았을 때, 특허의 LOF가 클수록 유망한 특허일 가능성이 크다고 볼 수 있지만, 오히려 그 반대임을 알 수 있었다. 따라서 기술의 메인스트림에 어느 정도 위치한 특허가 아니면 유망 특허가 아닐 가능성이 크다고 할 수 있다. 유망 특허의 규칙 2, 비유망 특허의 규칙 1에서 한 분야에 여러 기술이 생기다가 시간이 지나며 특정 기술만 사용되는 경우 같은 ‘coherence’에 관한 IKA, IKC의 경우도 도출된 시사점에 대한 근거가 된다. IKA는 같은 기간에 특허의 초록에서 쓰이는 의미 있는 키워드의 빈도인데, 해당 지표 또한 값이 클수록 대상 특허가 기술의 메인스트림에 위치함을 의미한다. 실험 결과에서는 IKA 1.5 기준으로 유망한 특허와 유망하지 않은 특허로 나누어지는 것을 확인하였다.

Ⅴ. 결 론

1. 연구의 요약 및 의의

본 연구는 특허의 구조적 부분인 서지학적 정보에 더해 비구조적 부분인 특허 텍스트 지표를 활용하여 유망 특허를 예측하는 프레임워크를 제안하였다. 제안한 프레임워크는 특허 데이터로부터 연구에서 정의한 지표를 추출하여 유망 기술의 특징에 따라 맵핑한 뒤, 머신러닝 기법을 활용하여 유망 특허를 예측하는 방법으로 구성된다. 프레임워크의 실효성을 검증하기 위해 사례 연구를 수행하였는데, 헬스케어 기술 분야의 특허를 분석하였다. 먼저 미국특허청(USPTO)에서 헬스케어 기술 관련 특허 고유번호를 수집하고, 이를 통해 PatentsView API 쿼리로 특허별 지표를 수집하였다. 수집한 지표들은 설계한 변수의 정의에 따라 전처리를 진행하였고, 특허의 유망성을 나타내는 1개의 종속변수와 특허의 서지학적 지표 13개, 특허의 텍스트 지표 3개로 구성된 16개의 설명변수를 얻었다. 각 지표의 특성에 따라 유망 기술의 특징인 ‘Radical Novelty’, ‘Fast Growth’, ‘Coherence’, ‘Prominent impact’, ‘Uncertainty and ambiguity’에 각각 맵핑하였다. 그리고 로지스틱 회귀, 의사결정나무, 랜덤 포레스트, LightGBM, XGBoost, AdaBoost, Gradient Boosting 및 Voting 분류 모델 총 10개의 모델을 활용하였고, 특허의 텍스트 지표를 고려한 모델과 그렇지 않은 모델로 학습 시나리오를 나누어 두 결과의 성능을 비교하였다. 그 결과 특허의 서지학적 지표와 텍스트 지표를 같이 고려한 모델이 그렇지 않은 모델보다 특허의 유망성을 예측하는 경우 더 나은 성능을 보였다. 연구에서 제안한 특허 텍스트 설명변수 중 특허의 초록에서 유의미한 기술 키워드의 빈도를 나타내는 IKA가 특허의 유망성에 유의미한 영향을 끼치는 것을 확인할 수 있었다. 또한, 의사결정나무를 통해 유망 특허, 유망하지 않은 특허의 규칙을 살펴보았다. 마지막으로 실험 결과들을 종합하여 유망한 특허는 기존의 통념과 달리 산업분야의 주류 기술을 사용하는 특허들이 될 가능성이 높다는 것을 확인하였다.

본 연구는 다음과 같은 의의를 가진다. 첫째, 유망 특허 예측 시 사용될 수 있는 특허의 내용을 반영하는 텍스트 변수를 제안하였다. 기존 연구에서는 주로 특허의 서지학적 정보를 주로 사용하였으며, 텍스트 정보를 사용하여도 임베딩 벡터로 사용하는 방향으로 연구가 진행되었다. 그러나 특허의 품질에 직접 영향을 주는 특허 초록의 내용, 청구항의 내용 등과 같은 특징을 반영하지 않았다. 본 연구에서 제안한 특허의 텍스트 변수는 유망 특허를 예측하는 경우, 예측 정확도 향상

과 영향력 측면에서 의의를 가진다. 둘째, 유망 기술 관점에서 유망 특허를 예측하였다. 이전 연구에서는 유망 특허 예측 시 유망 기술의 특징을 고려한 경우가 적었다. 유망 기술의 특징을 고려하였더라도 ‘fast growth’나 ‘coherence’같이 기술에 대한 지속적인 관찰이 필요한 특징을 배제하였다. 본 연구에서는 유망 기술의 특징 5가지 모두를 고려한 특허 변수를 사용하여 이러한 한계점을 개선하고자 하였다.

2. 연구의 한계점 및 추후 연구

본 연구는 다음과 같은 한계점도 존재한다. 첫째, 유망 클래스의 특허에 대한 예측 성능이 본 연구에서 기대한 수치보다 낮게 측정되었다. 이는 유망 기술의 다면적인 특징을 고려하지 않았기 때문이라고 생각된다. 따라서 신문, 논문 등의 다른 유형의 데이터로부터 추가로 지표를 추가하면 프레임워크 성능이 향상될 수 있을 것이다. 둘째, 본 연구에서는 특허의 텍스트 지표로 문서 내 키워드의 TF-IDF 점수, 문서의 LOF score 등의 지표를 활용하였다. 이때 TF-IDF 점수는 각 단어의 중요성을 판단하여, 해당 특허문서가 중요한 단어를 활용하는지 여부에 분석의 초점을 맞추고 있으며, LOF score는 해당 특허문서의 초록이 얼마나 참신한지에 대한 정도에 초점을 맞추고 있다. 두 지표 모두 해당 특허문서가 메인스트림에 있는 기술인지 여부를 판단하기 위해 활용된다. 그러나 본 연구에서는 희귀한 단어 또는 새로운 단어가 등장하는 경우는 고려하지 않고 있다. 예를 들어 기존 특허문서에서 한 번도 발생하지 않은 단어가 나타났는지 여부, 시점 t에서 새로운 단어의 출현 이후 시점에서 해당 단어가 급격히 증가하는 빈도 등 새로운 단어의 출현과 관련된 지표는 활용되지 않았다. 추후 연구에서는 기술의 신규성이라는 측면을 반영할 수 있는 다양한 텍스트 지표를 활용하면 더 의미 있는 연구가 될 것으로 생각된다. 셋째, 특허의 기술적 잠재력에 대한 기준으로 특허의 피인용 빈도를 이용하였다. 해당 지표는 특허 가치의 대용지표로써 장기적으로 기술적 가치로 쓰일 수 있지만, 특허의 가치는 해당 지표로 완전히 확인할 수는 없다. 추후 연구에서 특허의 라이센스 비용과 유지 기간 및 유지 비용 등의 지표 등을 분석에 반영하여 이를 보완할 수 있을 것이다. 넷째, 본 연구에서 사용한 특허 데이터 세트는 클래스 불균형 문제를 가지고 있다. 유망 클래스의 비율이 유망하지 않은 클래스보다 훨씬 낮기 때문에 오버샘플링 기법을 활용하였지만, 과적합 문제와 이상치에 민감하다는 단점이 발생할 수 있다. 이러한 단점은 더 많은 특허 데이터와 특허 품질을 측정하는 대용지표를 더 추가하는 추후 연구를 통해 해결할 수 있을 것으로 기대된다. 마지막으로 사례 연구의 기술 분야를 헬스케어 분야로 한정하였다는 한계점이 존재한다. 추후 연구에는 특허 데이터의 분야를 확장하여 연구를 진행할 필요성이 존재한다.

참고문헌

Anne, C., Mishra, A., Hoque, M. T., & Tu, S. (2018). Multiclass patent document classification. Artif. Intell. Res., 7(1), 1-14.

Bierly, P., & Chakrabarti, A. (1996). Determinants of technology cycle time in the US pharmaceutical industry'. R&D Management, 26(2), 115-126.

Breitzman, A., & Thomas, P. (2015). The Emerging Clusters Model: A tool for identifying emerging technologies across multiple patent systems. Research policy, 44(1), 195-205.

Callaert, J., Van Looy, B., Verbeek, A., Debackere, K., & Thijs, B. (2006). Traces of prior art: An analysis of non-patent references found in patent documents. Scientometrics, 69(1), 3-20.

Choi, J., & Hwang, Y. S. (2014). Patent keyword network analysis for improving technology development efficiency. Technological Forecasting and Social Change, 83, 170-182.

Chung, P., & Sohn, S. Y. (2020). Early detection of valuable patents using a deep learning model: Case of semiconductor industry. Technological Forecasting and Social Change, 158, 120146.

Cozzens, S., Gatchair, S., Kang, J., Kim, K. S., Lee, H. J., Ordóñez, G., & Porter, A. (2010). Emerging technologies: quantitative identification and measurement. Technology Analysis & Strategic Management, 22(3), 361-376.

Daim, T. U., Rueda, G., Martin, H., & Gerdsri, P. (2006). Forecasting emerging technologies: Use of bibliometrics and patent analysis. Technological forecasting and social change, 73(8), 981-1012.

Davis, J., & Goadrich, M. (2006, June). The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd international conference on Machine learning (pp. 233-240).

Day, G. S., & Schoemaker, P. J. (2000). Avoiding the pitfalls of emerging technologies. California management review, 42(2), 8-33.

Érdi, P., Makovi, K., Somogyvári, Z., Strandburg, K., Tobochnik, J., Volf, P., & Zalányi, L. (2013). Prediction of emerging technologies based on analysis of the US patent citation network. Scientometrics, 95(1), 225-242.

Ernst, H. (2003). Patent information for strategic technology management. World patent information, 25(3), 233-242.

Gerken, J. M., & Moehrle, M. G. (2012). A new instrument for technology monitoring: novelty in patents measured by semantic patent analysis. Scientometrics, 91(3), 645-670.

Geum, Y., Jeon, J., & Seol, H. (2013). Identifying technological opportunities using the novelty detection technique: A case of laser technology in semiconductor manufacturing. Technology Analysis & Strategic Management, 25(1), 1-22.

Girthana, K., & Swamynathan, S. (2018). Patent Document Clustering Using Dimensionality Reduction. In Progress in Advanced Computing and Intelligent Engineering (pp. 167-176). Springer, Singapore.

Harhoff, D., Scherer, F. M., & Vopel, K. (2003). Citations, family size, opposition and the value of patent rights. Research policy, 32(8), 1343-1363.

Haupt, R., Kloyer, M., & Lange, M. (2007). Patent indicators for the technology life cycle development. Research Policy, 36(3), 387-398.

Joung, J., & Kim, K. (2017). Monitoring emerging technologies for technology planning using technical keyword based analysis from patent data. Technological Forecasting and Social Change, 114, 281-292.

Ju, Y., & Sohn, S. Y. (2015). Identifying patterns in rare earth element patents based on text and data mining. Scientometrics, 102(1), 389-410.

Kayal, A. A., & Waters, R. C. (1999). An empirical evaluation of the technology cycle time indicator as a measure of the pace of technological progress in superconductor technology. IEEE Transactions on Engineering Management, 46(2), 127-131.

Kim, J., & Lee, C. (2017). Novelty-focused weak signal detection in futuristic data: Assessing the rarity and paradigm unrelatedness of signals. Technological Forecasting and Social Change, 120, 59-76.

Kim, J., & Lee, S. (2015). Patent databases for innovation studies: A comparative analysis of USPTO, EPO, JPO and KIPO. Technological Forecasting and Social Change, 92, 332-345.

Kwon, U., & Geum, Y. (2020). Identification of promising inventions considering the quality of knowledge accumulation: A machine learning approach. Scientometrics, 125(3), 1877-1897.

Kyebambe, M. N., Cheng, G., Huang, Y., He, C., & Zhang, Z. (2017). Forecasting emerging technologies: A supervised learning approach through patent analysis. Technological Forecasting and Social Change, 125, 236-244.

Lanjouw, J. O., & Schankerman, M. (2004). Patent quality and research productivity: Measuring innovation with multiple indicators. The Economic Journal, 114(495), 441-465.

Lanjouw, J., & Schankerman, M. (2001). Enforcing intellectual property rights. National Bureau of Economic Research.

Lee, C., Kim, J., Kwon, O., & Woo, H. G. (2016). Stochastic technology life cycle analysis using multiple patent indicators. Technological Forecasting and Social Change, 106, 53-64.

Lee, C., Kwon, O., Kim, M., & Kwon, D. (2018). Early identification of emerging technologies: A machine learning approach using multiple patent indicators. Technological Forecasting and Social Change, 127, 291-303.

Lerner, J. (1994). The importance of patent scope: an empirical analysis. The RAND Journal of Economics, 319-333.

Martin, B. R. (1995). Foresight in science and technology. Technology analysis & strategic management, 7(2), 139-168.

Narin, F., Noma, E., & Perry, R. (1987). Patents as indicators of corporate technological strength. Research policy, 16(2-4), 143-155.

Niemann, H., Moehrle, M. G., & Frischkorn, J. (2017). Use of a new patent text-mining and visualization method for identifying patenting patterns over time: Concept, method and test application. Technological Forecasting and Social Change, 115, 210-220.

OuYang, K., & Weng, C. S. (2011). A new comprehensive patent analysis approach for new product design in mechanical engineering. Technological Forecasting and Social Change, 78(7), 1183-1199.

Petruzzelli, A. M., Rotolo, D., & Albino, V. (2015). Determinants of patent citations in biotechnology: An analysis of patent influence across the industrial and organizational boundaries. Technological Forecasting and Social Change, 91, 208-221.

Porter, A. L., Roessner, J. D., Jin, X. Y., & Newman, N. C. (2002). Measuring national ‘emerging technology’ capabilities. Science and Public Policy, 29(3), 189-200.

Rotolo, D., Hicks, D., & Martin, B. R. (2015). What is an emerging technology?. Research policy, 44(10), 1827-1843.

San Kim, T., & Sohn, S. Y. (2020). Machine-learning-based deep semantic

analysis approach for forecasting new technology convergence. Technological Forecasting and Social Change, 157, 120095.

Schubert, E., Zimek, A., & Kriegel, H. P. (2014). Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection. Data mining and knowledge discovery, 28(1), 190-237.

Small, H., Boyack, K. W., & Klavans, R. (2014). Identifying emerging topics in science and technology. Research policy, 43(8), 1450-1467.

Squicciarini, M., Dernis, H., & Criscuolo, C. (2013). Measuring patent quality: Indicators of technological and economic value, OECD Science, Technology and Industry Working Papers, No. 2013/03, OECD Publishing

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

Abstract

Predicting Emerging Technologies through Patent Data: Based on Characteristics of Emerging Technologies

Jeon, Woo Jin
(Supervisor Geum, Young Jung)
Dept. of Data Science
Graduate School
Seoul National University of Science and Technology

Identifying the drivers of technological changes leading to future social changes and R&D on core technologies are considered important factors for improving technological and industrial competitiveness at the national and corporate level, and emerging technologies are of great interest to industries and government officials. However, due to the rapid pace of technological development in recent years, companies are having difficulty selecting suitable technologies for investment. As a solution to this, various studies have been conducted to predict emerging technologies using patent data. However, in previous studies, emerging technologies were predicted mainly by focusing on the structural part of patents, and the research was conducted without considering the characteristics of promising technologies. However, when evaluating the quality of a patent, it is necessary to consider the information of unstructured part of the patent, and at the same time, it is also necessary to consider the multi-faceted features of technology emerging technology. Therefore, this study proposes a systematic method for predicting emerging patent in consideration of the unstructured part of the patent and the characteristics of the promising technology. This study extracted bibliographic indicators from patent data based on previous studies using the structural part of the patent, proposed text-related indicators, which are unstructured parts of the patent, mapped them according to five characteristics of the promising technology, and predicted the emerging patent. 14,992 of patent data related to healthcare technology were used which are collected through the PatentView API. Descriptive variables defined in advance

from patent data were extracted and calculated, and the prospect of patents was predicted through machine learning. As a result of the experiment, it was confirmed that the patent text indicator proposed in this study, frequency of meaningful keywords used in the abstract of patents during the same period, is an important factor in determining the emerging patents. The framework for predicting emerging patent presented in this study is expected to help identify patents and technologies suitable for companies to invest in.

감사의 글

먼저 2년 동안 저를 지도해 주신 금영정 교수님께 감사의 인사를 드리고 싶습니다. 학부 때부터 진심 어린 조언과 격려를 해주신 교수님 덕분에 데이터 사이언스 분야로 진로 방향을 정할 수 있었습니다. 많이 부족했던 저를 항상 기다려주시고, 올바른 연구자가 될 수 있도록 도와주셔서 무사히 졸업할 수 있게 되었습니다. 저절로 존경하게 되는 교수님의 인품을 본받아 큰마음을 가진 사람이 될 수 있도록 노력하겠습니다. 교수님의 제자가 될 수 있어 영광이었습니다.

제가 하는 모든 결정에 고민을 함께 해주시고 힘을 실어주신 가족들에게 감사드립니다. 항상 제가 연구하는 분야에 대해 관심을 가져주시고 든든한 버팀목이 되어주신 아빠, 궁금증이 많으시지만 재촉하지 않고 묵묵히 기다려주신 엄마에게 항상 감사하고 사랑한다는 말을 전하고 싶습니다. 그리고 나이가 들며 힘이 되어준 누나는 즐거운 신혼 생활이 됐으면 좋겠습니다.

대학원 동기들에게도 정말 고맙다고 말하고 싶습니다. 먼저 학부부터 같은 연구실까지 함께 연구하며 정신적 버팀목이 되어준 재웅, 같은 연구실을 공유하며 ‘조진웅’의 멤버로서 저의 짓궂은 장난을 잘 받아줬던 인조, 묵묵히 힘이 되어준 호현이 형. 연구하며 힘든 순간들을 덕분에 버틸 수 있어서 감사했습니다. 또한, 서비스혁신 연구실에서 함께 연구하며 든든했던 소희누나, 성인군자의 표본이셨던 민규형, 코드가 잘 맞아 함께 대화하면 즐거웠던 병민, 철없는 오빠들과 수준 맞춰 잘 놀아주고 연구를 도와줬던 예빈, 다방면으로 능통해 듣는 즐거움이 있던 영준이형에게 고마웠다고 말하고 싶습니다. 연구하며 별일 없을 때도 기꺼이 연락 주시면 다들 술 한잔 사드리겠습니다.

서류 작업에 대해 모르는 게 있을 때마다 질문하면 친절히 알려주셨던 송미자 선생님과 노민영 선생님에게도 감사 인사를 드리고 싶습니다.

마지막으로 논문을 심사해주신 조남욱 교수님과 이영훈 교수님께 감사의 인사를 전하며, 데이터 사이언티스트로 성장할 수 있게 도와주신 모든 데이터사이언스학과 교수님들도 감사드립니다. 2년간의 가르침을 바탕으로 앞으로 훌륭한 연구자가 그리고 올바른 사람이 될 수 있도록 노력하겠습니다. 감사합니다.