Journal of the Korean Institute of Industrial Engineers
Vol. 50, No. 1, pp. 47-63, February 2024.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2024.50.1.047

특허-상표 연계 비즈니스 인텔리전스를 위한 텍스트 분석 기반의 비즈니스 영역 식별

윤주호 · 김병훈†

한양대학교 산업경영공학과

Text Analytics-based Business Area Identification for Patent-Trademark Linkage Business Intelligence

Juho Yoon · Byunghoon Kim

Department of Industrial and Management Engineering, Hanyang University

This study presents a novel approach for identifying new business opportunities by analyzing the linkage between patents and trademarks leveraging text analytics. Initially, we utilize topic modeling to analyze the descriptions of goods and services in trademarks, with a particular focus on trademarks that do not share similar group codes. Using the Latent Dirichlet Allocation (LDA) model, the descriptions in the trademarks are segmented into multiple business groups based on similarities. Subsequently, we define business areas by measuring their similarity to the industry classifications represented by the Standard Industrial Classification (SIC) system. To this end, we propose a novel weighted cosine similarity. Leveraging the proposed similarity, we align each patent with one of the predefined business groups extracted from the trademark data. Based on this approach, we can identify business areas closely related to the technological capabilities of tech-based firms. In the case study, we showed that business areas are identified through the alignment between the customized goods and service groups and SIC from trademark data of global technology-based firms.

Keywords: Business Intelligence, Patent-Trademark Linkage Data, Similar Groups, Technology-Based Firms, Text Analytics

1. 서 론

비즈니스 인텔리전스(business intelligence)는 기업 내외부의 다양한 정보로부터 수집된 데이터를 분석하여 효과적이고 전략적인 경영 의사결정을 지원하기 위한 방법 및 시스템을 통칭한다(Davenport, 2006; Chen *et al.*, 2012). 비즈니스 인텔리전스를 위한 소재로는 기술자료, 재무자료, 운영자료, 시장자료 등이 있으며 특허나 상표와 같은 지식재산권(intellectural property, 이하 IP)과 같은 무형자산도 기술개발과 사업 활동에 활용될 수 있다.

특히 특허와 상표 분석은 해당 기업의 기술적인 내용, 기업이 시장에 제공하고자 하는 상품과 서비스를 분석하는 데 유용하며, 이러한 분석은 비즈니스 경쟁 분석, 비즈니스 다각화 분석 등과 관련된 인텔리전스를 제공한다(Doern, 1999; Sandner, 2008). 실제로 특허 데이터를 기반으로 기업 의사결정을 지원하기 위해 기술지식흐름분석(Yoon *et al.*, 2015; Song *et al.*, 2017), 기술로드맵핑(Jeong and Yoon, 2015), 기술기회도출(Park and Yoon, 2017), 융합동향분석(Caviggioli, 2016), 기술예측(Yoon *et al.*, 2018)과 같은 다양한 연구들이 진행되었다. 그러나 특허 기반의 분석은 기업의 상품이나 서비스와 같은 사업영역을 포괄적으로 분석하는데 일부 한계가 있다. 상표 기반의 분석은 기업이 시장에 제공하고자 하는 상품과 서비스로 인식하고 관련된 분석을

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2022R1F1A1063273).
† 연락저자 : 김병훈 교수, 15588 경기도 안산시 상록구 한양대학로 55 5공학관 510호 산업경영공학과, Tel : 031-400-5269, E-mail : byungkim@hanyang.ac.kr
2023년 10월 25일 접수; 2023년 11월 15일 수정본 접수; 2023년 11월 21일 게재 확정.

지원하는 고유의 장점을 가지고 있다. 그러나 특허에 비해 독립적으로 분석하기에 매력적인 하위 구성 요소들이 부족하고(Yoon *et al.*, 2019), 수집할 수 있는 채널이 제한적이다(Ko, 2022). 또 상표는 10년마다 갱신할 수 있는 반영구적인 존속기간을 가지며(Chudnovsky, 1983), 특허 출원 이후 해당 기술을 적용한 신제품과 서비스를 출시하는 경향이 있다(Ferreira and Godinho, 2011). 특히 미국의 경우 사용주의 원칙에 따라 실제 사용증명서(statement of use)를 제출해야 상표로 등록되는 등의 이유로 상표를 활용한 비즈니스 분석 연구는 국내외에서 부족한 실정이다(Jeong, 2010).

실제 비즈니스 환경에서의 경쟁은 유사한 제품과 서비스를 가진 기업 간에 발생한다. 따라서 특허나 상표 데이터를 단독으로 사용하기보다는 이를 연계(linkage)하여 분석하는 것이 중요하다. 즉, 기술 기반 기업(technology-based firms)이 시장에 제공하는 실제 제품과 서비스에 대한 정보를 고려하고, 비즈니스 영역(business area)에서 활용하는 기술과 활용하고 있지 않은 기술을 구분함으로써 효과적인 비즈니스 분석을 가능케 한다(Ko, 2022). 그러나 속지주의 원칙(territorial principle)을 따르는 IP의 특성상 국가별 상표 제도에 따른 차이로 비즈니스 인텔리전스 소재로써 활용하는 데 한계가 있다(Moon, 2015).

한국의 상표 제도는 상품은 G로, 서비스업은 S로 시작되며 총 5-7자리의 문자와 숫자로 구성된 유사군 코드(similar group code)를 활용한다. 다양한 상품들 또는 서비스들을 유형화(grouping)하여 상표를 분류하고, 특허청에서 정의한 지정상품(goods and services)을 사용하도록 권고하고 있어 상표 데이터가 담고 있는 내용이 상대적으로 구체적이다. 반면에 미국의 상표 제도는 상표자, 날짜 정보, NICE Class와 텍스트로 기재된 제품과 서비스에 대한 간략한 설명이 전부이다. 미국특허청(united states patent and trademark office, 이하 USPTO)에서는 상표 전자 신청 시스템(trademark electronic application system, 이하 TEAS)을 통해 TEAS Standard와 TEAS Plus 두 가지 유형으로 상표 출원이 가능하고, 출원 유형에 따라 customized goods and service를 표기할 수도 있으며 이는 보호범위와 직결되기 때문에 <Table 1>과 같이 실제 제품과 관련이 없는 포괄적인 지정상품을 사용하기도 한다(Choi and Yoon, 2021).

Ko *et al.*(2020)은 국내 상표의 지정상품명을 기술 기반 기업의 비즈니스 영역으로 간주하고, 임베딩 모델을 활용하여 <Table 2>와 같이 기술 기반 기업의 기술과 비즈니스 영역 간

Table 1. Trademark SAMSUNG GALAXY WATCH in KIPRIS and USPTO

Country	Class / Similar group code	Goods and Service
KIPRIS (KOREA)	09 / G390702	Mobile digital telecommunication devices
	09 / G390802	Software
	09 / G390701, G390702, G390803	Wrist smart phone
	09 / G390701, G390702, G390803	Smart glasses
	09 / G390701, G390702, G390803	A watch that transmits data to a smartphone
	09 / G390701, G390702, G390803	Watch-shaped smartphone
	09 / G390702	Watch-shaped smartphone band
	09 / G390102	Watch-shaped smartphone charger
	09 / G390702	Watch-shaped smartphone strap
	09 / G3905	Watch-shaped cable for smartphone
	09 / G390701, G390702, G390803	Smartphone functions in wearable watch form
	09 / G390803	Wearable computer peripherals
	09 / G390701, G390702, G390803	Wearable computer
USPTO (USA)	NICE Class	(Based on Use in Commerce) Wrist-mounted smartphones; Smart watches that communicate data to smartphone; Smartphones in the shape of a watch; Wearable digital electronic devices incorporating smartphones in the shape of a watch in the nature of smart watches; Wearable digital electronic devices, comprised primarily of smart watchbands that communicate data to smartphones and comprised primarily of software and display screens for viewing, sending and receiving texts, emails, data and information to smartphones all in the nature of smart watches; Smart watches that communicate data to smartphones, through Internet websites and other computer and electronic communication networks; Smart watches incorporating cameras and MP3 players, ….
	9, 14	

의 연계를 수행하였다. 그러나 사전 실험을 통해 동일한 방법으로 미국 특허와 상표를 가진 기술 기반 기업의 기술과 비즈니스 영역을 연계한 결과 <Table 3>과 같이 기술과 상표 지정상품의 유사도를 계산할 수는 있지만, 한국 상표의 유사군 코드 수준과 같은 구체적인 비즈니스 영역은 식별할 수 없다는 한계가 있다. 따라서 본 연구에서는 상표 데이터의 지정상품이 구체적으로 유형화되지 않은 경우, 그룹화된 토픽을 통해 산업분류를 파악함으로써 비즈니스 영역을 식별하는 방법을 제안한다. 이를 위해서 먼저, 잠재 디리클레 할당(latent dirich allocation, 이하 LDA) 모델을 이용하여 상표의 지정상품 텍스트를 분석하고, 이를 유사군으로 표상(similar group representation)한다. 다음으로 표상된 유사군과 표준산업분류코드(standard industrial classification, 이하 SIC)의 유사도 비교를 통해 비즈니스 영역을 식별한다. 이때 가중된 코사인 유사도(weighted cosine similarity)를 사용하여 보다 정교한 비즈니스 영역이 매칭되도록 한다. 마지막으로 임베딩 모델을 활용하여 기술 기반 기업의 특허 데이터와 식별된 비즈니스 영역의 의미론적 분석을 통해 기술-비즈니스 영역을 연계한다.

본 연구는 학술적 관점에서 특허-상표를 연계한 비즈니스 인텔리전스를 위해 상표 데이터의 지정상품과 산업분류코드의 매칭을 통해 구체적인 비즈니스 영역을 식별한 초기 연구라는 점에 의의가 있다. 특허 데이터의 기술 정보와 상표 데이터에 있는 제품과 서비스에 관한 정보를 정리하고, 이들 간의 관계를 분석함으로써 기술적 관점에 국한되지 않고 제품과 서비스 관점도 고려하여 분석한다. 이는 특허와 다양한 외부 데이터를 결합한 다양한 후속 연구의 Best Practice가 될 것으로 기대된다. 실무적 관점에서는 기술 기반 기업들의 실제 비즈니스 경쟁 환경을 고려한 정량적 의사결정 지원이라는 점에 의의가 있다. 비전문가들도 데이터에 기반하여 비즈니스 분석을 가능케 함으로써 인적, 물적 자원이 부족한 글로벌 경쟁 환경 속에서 기술 기반 창업기업, 중소기업들의 Cold Start를 적시에 효율적으로 지원할 것으로 사료된다.

본 논문은 다음과 같이 구성된다. 제2장에서는 데이터 기반 비즈니스 인텔리전스 연구를 고찰하고 제3장에서는 제안 방법론을 설명한다. 제4장에서는 제안 방법론의 실시 예를 보이고, 제5장에서 본 연구를 정리한다.

Table 2. Example of Linked KIPRIS Trademark Goods and Service by KIPRIS Patent

Application Number Title (ENG)	Source (Sim)	Business area (ENG)
1020150002451 통신단말기 인증처리 시스템, 통신단말기, 단말기 인증서버 및 그 인증처리방법 Communication Terminal Certification Processing System, Communication Terminal, Server and Certification Processing Method (Dong-han, 2017)	Title (0.7983)	개인휴대통신단말기 (Personal handheld communication terminals)
	Title (0.8180)	GPS단말기 (GPS terminals)
	Title (0.6458)	휴대폰단말기 (mobile terminals)
	Abstract (0.6102)	서버관리업 (Server administration)

Table 3. Example of Linked USPTO Trademark Goods and Service by USPTO Patent

Application Number Title (ENG)	Source (Sim)	Goods and Service
17345674 Systems and computerized methods for balancing inventor (Yang *et al.*, 2022)	Title (0.7164)	Gallium; sensitized paper, namely, photographic paper and chemical test paper; china clay; flour for industrial purposes; industrial chemicals; adhesives for industrial purposes; ceramic glazings in the nature of a dry chemical preparation for use in the manufacture of ceramics; ···, action skill and arcade games; toys being playthings, namely, action figure toys, bath toys and bendable toys; fairground ride apparatus in the nature of amusement park rides; ornaments for Christmas trees, except illumination articles and confectionery, Maintenance and repair of cameras for CCTVs; repair of DC power generators being alternators; furniture restoration; repair of bags; rental of gas or air compressors; ···.
	Abstract (0.7378)	

2. 관련 연구

본 장에서는 이론적 관점에서 데이터 기반의 비즈니스 인텔리전스 연구를 대관하고, 방법론적 관점에서 텍스트 분석을 활용한 비즈니스 인텔리전스 연구를 세찰함으로써 본 논문의 연구범위를 명확히 한다.

2.1 데이터 기반 비즈니스 인텔리전스

최근 급변하는 비즈니스 환경에서 전문가의 경험과 지식만으로 경영 환경과 경쟁 동향을 지속적으로 파악하는 것에 한계가 생기면서, 데이터 기반 비즈니스 분석은 기술과 비즈니스에 대한 경쟁 정보를 결정하는 합리적인 접근 방식으로 주목받고 있다(Shibata *et al.*, 2008). 특히 IP로 대표되는 두 축인 특허와 상표 데이터는 독창적인 지식을 포함하는 동시에 경제적 가치를 지닌 자원으로 많은 연구에서 비즈니스 인텔리전스 소재로써 활용되었다.

Oltra and Saint Jean(2009)은 특허 데이터를 사용하여 저공해 자동차 분야에서 경쟁 기술을 식별하였고, Hao *et al.*(2010)은 특허 데이터를 활용하여 RFID와 관련된 기업을 식별하고 모니터링하였다. Lee and Yang(2015)는 특허에서 추출한 시장 공통성과 자원 유사성을 기반으로 경쟁자를 분석하였으며, Luo and Su(2015)는 기술발전의 추이와 이동통신산업의 경쟁상황을 조사하기 위해 기술 분야의 경쟁 동향을 분석하는 특허 기반의 정량적인 연구를 수행하였다. Pargaonkar(2016)은 특허 환경과 전략적 의도에 따른 경쟁 지능에 대한 프레임워크를 제안하였고, Kim *et al.*(2016)은 특정 기술 영역에 있는 기업의 특허 정보를 통해 신기술 영역의 출현을 탐색했다. Jeong and Yoon(2017)은 증강 현실 기술 분야에서 활용하기 위한 특허 기반의 경쟁 지능 분석 프레임워크를 제안하였으며, Wang *et al.*(2018)은 서지정보, 특허 분석, 기술로드맵 방법을 활용하여 새로운 나노발전기 기술을 식별하는 연구를 수행했다. Seip *et al.*(2018)은 상표를 기반으로 기술 기반 기업의 비즈니스를 분석하였고, Barroso *et al.*(2019)은 상표 데이터를 활용하여 제품 성능을 분석하였다. Ko *et al.*(2020)은 국내 기술 기반 기업들의 특허 데이터를 해당 기업의 기술로, 상표 데이터의 유사군 코드는 비즈니스 영역으로 정의한 뒤 임베딩 모델로 단어 간 유사도 계산을 통해 기술-비즈니스 영역을 연계한 비즈니스 인텔리전스를 수행하였다. 이를 통해 특허나 상표를 단독으로 사용하여 실제 비즈니스 환경을 반영하지 못한 기존 연구들의 한계를 극복하고, 나아가 기술 기반 기업이 지정한 제품과 서비스들 중에서 해당 기업이 보유한 기술을 활용하여 실제 비즈니스를 운영하는 비즈니스 영역을 필터링하였다.

기존 연구를 종합하여 볼 때 특허 데이터를 기반으로 한 연구들은 기술분야, 경쟁기업, 경쟁기술 차원에서 비즈니스 분석이 가능하지만, 실제 경쟁이 이루어지는 제품과 서비스 차원(product & service levels)의 비즈니스 환경은 고려하지 못하는 한계가 있다. 또한 특허와 상표 데이터를 모두 활용한 경우라도 상표 지정상품의 유사군이 구체적으로 지정되어 있지 않으면 비즈니스 영역을 식별하기 어렵고, 동일한 제품과 서비스이지만 텍스트의 미세한 차이로 서로 다른 비즈니스 영역으로 식별되는 문제도 있다. 이에 본 연구에서는 특허-상표를 연계하여 비즈니스 인텔리전스 소재로 활용하기 위해 상표의 지정상품 텍스트 분석을 통해 비즈니스 영역을 식별하는 방법을 제안한다.

2.2 토픽 모델링 기반의 비즈니스 영역 식별

본 절에서는 방법론적 관점에서 토픽 모델링을 활용한 비즈니스 인텔리전스 연구를 살펴본다. 토픽 모델링은 텍스트 텍스트의 문서 집합을 구성하는 단어를 분석함으로써, 잠재적인 주제를 도출하는 확률 기반 모형을 일컫는다(Blei and Lafferty, 2009). 토픽 모델링을 위한 기법으로는 LSA(latent semantic analysis), pLSA(probabilistic latent semeantic analysis), LDA 등이 있다. 특히 LDA는 문서의 수가 증가함에 따라 발생하는 과적합(overfitting)에 강건하여 우수한 문서 분류 성능뿐만 아니라, 문서 수준에서 확률 모델을 제공하여 결과 해석이 용이하다는 장점이 있다(Lu *et al.*, 2011). 비즈니스 인텔리전스에서는 이러한 토픽 모델링을 활용하여 동향 분석, 기술경쟁 정보 분석 등의 목적으로 활용된 바 있다(Kim *et al.*, 2016).

Oh *et al.*(2017)은 증강현실 기술의 동향을 분석하기 위해 미국 특허를 대상으로 LDA를 활용하여 세부기술 요소를 도출하고 추세 분석에 활용하였고, Jeong *et al.*(2019)는 특정 스마트폰에 대한 고객들의 리뷰 텍스트에 LDA를 활용하여 고객들의 요구 사항, 불만 등을 내포한 주제들을 산출하여 분석에 사용하였다. 뿐만 아니라 Lee *et al.*(2018)은 특허 데이터의 구성요소인 '해결하고자 하는 문제'와 '해결 방법' 텍스트에 각각 LDA를 적용하여 공통의 문제와 해결방법을 주제의 형태로 파악하여 기술 경쟁 동향 분석에 활용한 바 있다.

본 연구는 이에 착안하여 <Table 1>의 미국 상표 데이터와 같이 지정상품이 구체적으로 지정되어 있지 않은 경우, LDA를 활용하여 주제를 도출함으로써 유사군을 표상하고 이를 통해 비즈니스 영역을 식별함으로써 기존 연구의 한계를 극복하고자 한다.

3. 제안 방법론

상표 지정상품의 유사군이 없는 경우 비즈니스 영역을 식별하고, 기술-비즈니스 영역을 연계하는 과정은 <Figure 1>과 같이 3단계로 진행된다. 첫 번째 단계에서는 LDA 토픽 모델링을 통해 유사군을 표상하고, 두 번째 단계에서는 표상된 유사군과 SIC의 유사도를 계산하여 비즈니스 영역을 식별한다. 마지막 단계에서는 임베딩을 기반으로 기술-비즈니스 영역의 의미론적 유사도를 비교하여 특허-상표 데이터를 연계한다. 자세한

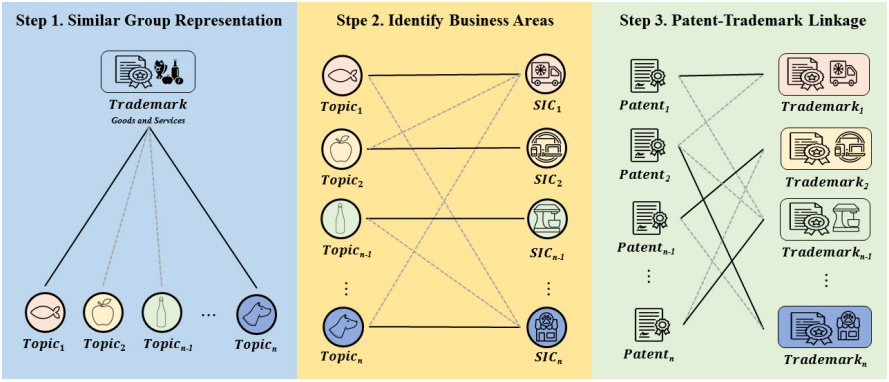


Figure 1. Framework of the Proposed Methodology

학습 과정은 다음 절에서 차례로 설명한다.

3.1 상표 지정상품 텍스트 전처리

제안 방법론을 수행하기에 앞서 상표 지정상품 데이터가 텍스트 형태이므로 이를 전처리함으로써 벡터화하는 과정이 필요하다. 우선 사전에 정의된 불용어, 구두점, 특수문자를 제거하고, 모든 문자를 소문자로 변환한다. 이후 텍스트를 단어 단위로 토큰화하고, 각 단어들의 표제어를 추출함으로써 텍스트를 정제하고 통일성 있게 한다. 이렇게 정제된 텍스트는 TF-IDF(term frequency-inverse document frequency)를 활용하여 식 (1)과 같이 상표 j에 대한 단어 i의 가중치($W_{i,j}$)로 벡터화한다(Salton and Mcgill, 1986).

$$W_{i,j} = tf_{i,j} \times \log\left(\frac{D}{df_i}\right) \tag{1}$$

여기서 $tf_{i,j}$는 상표 j에 나타나는 단어 i의 빈도를 의미하고, df_i는 단어 i를 포함하는 상표의 수를 나타낸다. D는 전체 상표의 수를 나타낸다.

3.2 잠재 디리클레 할당 기반 유사군 표상

전처리 과정에서 생성된 문서 벡터를 이용하여 LDA 기반의 토픽 모델링을 활용하여 식 (2)와 같이 상표 데이터로부터 토픽 주제어를 추출한다. 토픽 모델링은 상표 d에 단어 w가 나타날 조건부 확률 $p(w \mid d)$을 독립 다항 분포로 가정하고 은닉 변수(hidden variable)인 토픽 z를 추가하여 토픽-단어 간 조건부 확률 $p(w \mid z)$과 상표-토픽 간 조건부 확률 $p(z \mid d)$로 분해하여 $p(w \mid d)$를 추정하는 방법론이다(Jeong, 2015).

$$p(w \mid d) = \sum_{z} p(w \mid z)p(z \mid d) \tag{2}$$

본 연구에서는 $p(w \mid d)$를 추정하기 위해서 latent semantic indexing(Hofmann, 1999)의 사전 확률을 디리클레 분포로 가정한 LDA를 사용한다(Blei, 2012). LDA 토픽 모델링은 D개의 상표에서 상표별 토픽비율 (θ_d)과 K개의 토픽에서 토픽별 단어분포(Φ_k)는 각각 식 (3)과 식 (4)에 의해 결정된다.

$$\theta_d \sim Dirichlet(\alpha), d \in \{1, ..., D\} \tag{3}$$

$$\Phi_k \sim Dirichlet(\beta), k \in \{1, ..., K\} \tag{4}$$

여기서 α와 β는 각각 θ_d, Φ_k에 대한 디리클레 하이퍼파라미터(dirichlet hyperparameter)로 이를 이용하여 N개의 단어 중 d번째 상표의 i번째 단어는 식 (5)과 식 (6)에 의해 토픽($z_{d,i}$)과 단어($w_{d,i}$)가 결정되고, 이를 도식화하면 <Figure 2>와 같다.

$$z_{d,i} \sim \ Multinomial(\theta_d), i \in \{1, ..., N\} \tag{5}$$

$$w_{d,i} \sim Multinomial(\Phi_{z_{d,i}}) \tag{6}$$

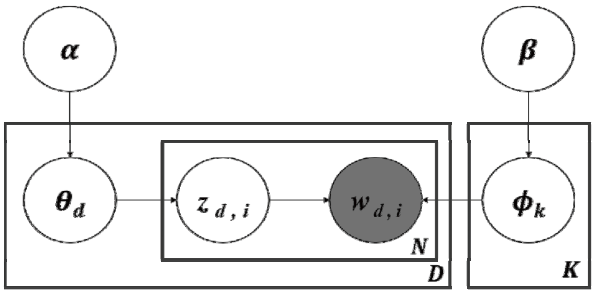
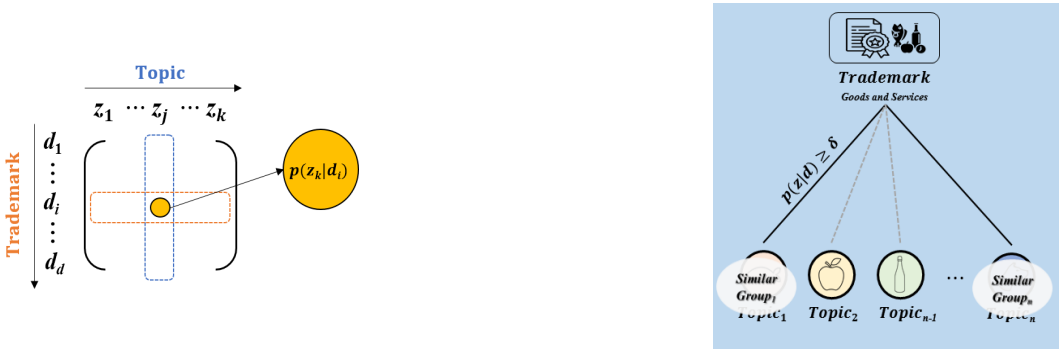


Figure 2. Graphical Model of LDA

이처럼 전체 상표 데이터에 LDA 토픽 모델링을 통해서 <Figure 3>의 (a)와 같이 각 상표의 토픽 분포를 나타내는 상표-토픽 매트릭스가 생성되고, 이때 $p(z \mid d)$값에 임곗값(threshold)을 주어 일정 수준 이상(δ)인 경우 <Figure 3>의 (b)와 같이 해당 상표 지정상품의 유사군으로 활용한다.

3.3 가중된 코사인 유사도 기반 비즈니스 영역 식별

한편 표상된 유사군은 주제별로 클러스터링 된 단어의 분포



(a) Trademark-Topic Matrix (b) Similar Group Representation

Figure 3. Similar Group Representation based on LDA

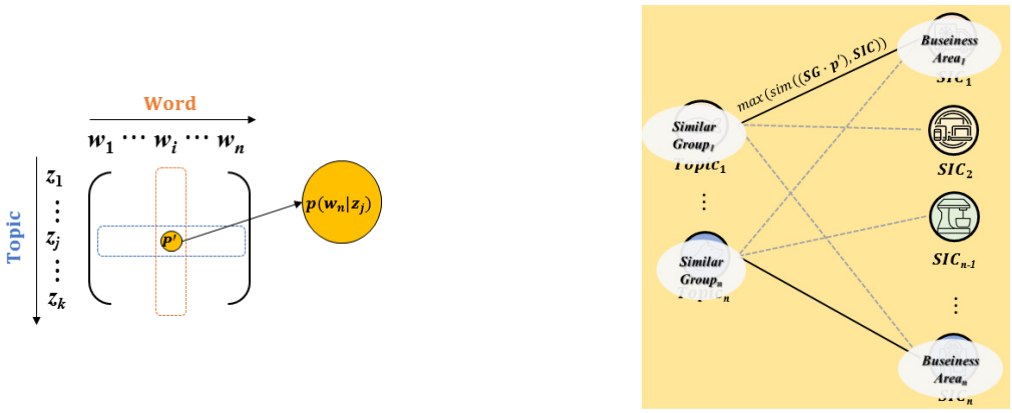
이므로, 그 자체를 비즈니스 영역으로 보기에는 어려움이 있다. 따라서 본 절에서는 표상된 유사군과 SIC의 유사도를 기반으로 비즈니스 영역을 식별하기로 한다.

1940년대 미 정부기관인 OMB(office of management and budget)에 의해 개발된 SIC는 미국 내에서 피분류기업이 속한 산업을 확인하기 위해 가장 흔히 쓰이는 분류체계이다. 그 이유는 SIC가 해당 기업의 경제활동을 가장 잘 설명하여주기 때문이다(Shine, 2005). SIC 분류체계는 네 자리 숫자와 레이블(description)로 구성된다. 우선 전체 산업이 11개 부문(division)으로 나누어져 있고, 각 부문은 다시 주요 집단(major group), 산업 집단(industry group), 산업(industry) 순의 계층적 구조로 나누어진다(Shin and Park, 2006). 예를 들어 <Table 4>에서와 같이 제조 산업을 가리키기 위한 SIC 코드는 SIC 2047이라는 네 자리 숫자로 표시가 되며, 'Dog and Cat Food'라는 레이블이 부착되어 있다.

이같은 SIC 레이블을 상표 지정상품과 동일한 방식으로 텍스트 전처리를 진행하고, 전처리 된 단어들은 사전에 학습된 fastText(Bojanowski *et al.*, 2017) 모델로 임베딩한다. 여기서 Ko *et al.*(2020)과 동일한 방식으로 SIC 레이블에 상표의 지정상품이 등장하는지 여부를 확인하기 위해서 SIC 레이블로 임베딩 모델을 학습한다. 자세한 임베딩 과정은 다음 절에서 다루도록 한다. 임베딩된 유사군 벡터와 SIC 벡터간 코사인 유사도 계산을 통해 비즈니스 영역을 식별한다. 이때 유사군을 구성하는 단어별 확률값 즉, <Figure 4>의 (a)에 해당하는 단어별 가중치(p')를 고려하여 식 (7)과 같이 가중된 유사도를 계산함으로써 보다 정교한 매칭이 되도록 한다.

Table 4. Example of SIC Classification

Industry Title	Division	Major Group	Industry Group	SIC	Description
			⋮		
Manufacturing	D	20	204	2047	Dog and Cat Food
			⋮		
Service	I	72	725	7251	Shoe Repair Shops and Shoeshine Parlors



(a) Topic-Word Matrix (b) Identify Business Areas

Figure 4. Business Area Identification based on Weighted Cosine Similarity

$$BA = \arg\max_{i}\left(sim\left(\left(SG_i \bullet p'_{i,SG_i}\right), SIC_j\right)\right) \quad (7)$$

비즈니스 영역 BA는 유사군 SG_i와 j번째 SIC 레이블 SIC_j 사이의 최대 유사도를 기반으로 식별된다. 여기서 SG_i는 문서 d에 대하여 $p(z \mid d) \geq \delta$를 만족하는 i번째 유사군을 나타내며, p'_{i,SG_i}는 이 유사군을 구성하는 각 단어 i의 확률값을 의미한다. SIC_j는 j번째 SIC 레이블을 대표하는 키워드 집합이다. $sim\left(\left(SG_i \bullet p'_{i,SG_i}\right), SIC_j\right)$는 SG_i와 SIC_j 사이의 가중된 코사인 유사도를 계산하며, 이는 SG_i 내 단어 확률 p'_{i,SG_i}와 SIC_j 키워드 집합 간의 연관성을 의미한다.

상기 과정을 통해 <Figure 4>의 (b)와 같이 표상된 유사군을 SIC와 가중된 코사인 유사도 계산을 통해 비즈니스 영역으로 식별이 가능해진다. 다음 절에서는 기술-비즈니스 영역 연계를 통한 비즈니스 분석을 위해 임베딩 기반의 특허-상표 데이터의 연결 방법을 다루도록 한다.

3.4 임베딩 기반 특허-상표 데이터 연결

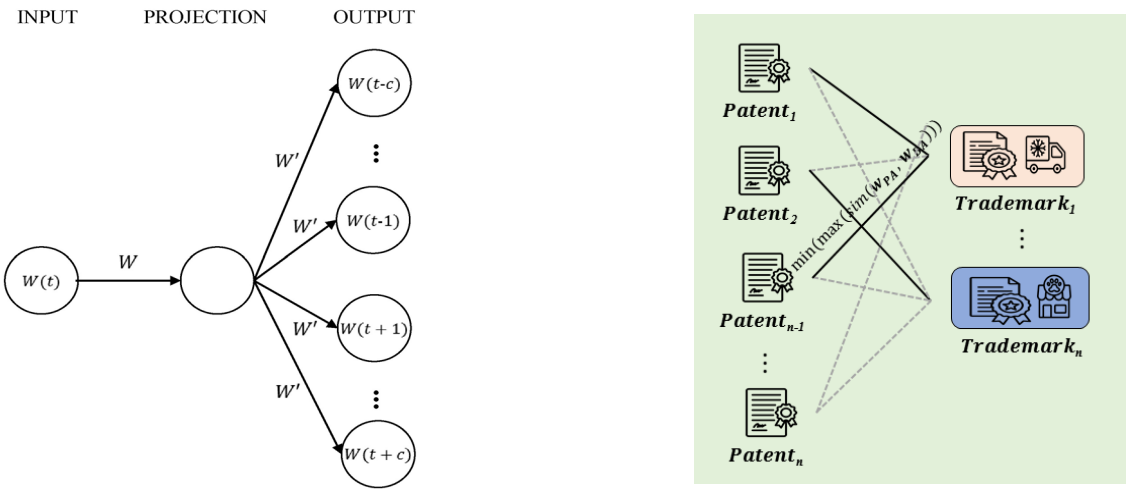
특허와 상표 데이터간 링크를 구축하는 가장 간단한 방법은 제목과 초록을 나타내는 특허 텍스트에 상표 데이터의 지정상품이 등장하는지 여부를 확인하는 것이다(Ko *et al.*, 2020). 이를 위해서 각 단어를 연속된 공간상의 실수로 변환하는 분산표상(distributed reprsentation)이 필요하다. 분산표상은 단어의 의미적(semantic), 구문적(syntactic) 정보를 반영한 의미론적 분석이 가능하다는 장점이 있다(Park *et al.*, 2021). 이러한 분산표상 방식은 3.1절의 TF-IDF와 같은 카운트 기반 방법과 신경망 모델을 적용한 예측 기반 방법으로 나뉜다.

신경망 구조를 이용한 대표적인 모델인 Word2Vec은 CBOW (continuous bag-of-words)와 Skip-gram 방식으로 단어를 예측하여 임베딩을 수행한다. CBOW 모델은 문장이나 문서에서 윈도우 크기만큼의 주변 단어가 주어질 때 중심 단어를 잘 예측하는 단어 표상을 찾도록 학습한다. 이때 주변 단어를 One-hot vector의 형태로 입력하면 가중치 행렬 W를 통해 원하는 차원으로 사영되고 가중치 행렬 W'을 통해 중심 단어가 출력된다. 모델은 중심 단어가 나올 확률을 최대로 하는 가중치 행렬 W와 W'를 학습하게 되며 최종적으로 W가 임베딩 행렬이 된다. 반면 Skip-gram 모델은 중심 단어가 주어질 때 주변 단어를 예측하는 방법으로 입력과 출력이 CBOW 방식과 반대로 작동한다. <Figure 5>의 (a)는 윈도우 크기(window size)가 c일 때 Skip-gram의 구조를 나타낸 것이다(Mikolov *et al.*, 2013).

본 연구에서는 특허-상표 데이터 연결 기반의 비즈니스 인텔리전스 초기 연구를 확장한다는 점에서 Ko *et al.*(2020)과 동일한 방식으로 Skip-gram 기반의 분산표상 방식인 fastText 모델을 활용한다. fastText 모델은 사전 학습된 모델을 사용하거나 도메인별 텍스트 데이터를 기반으로 학습시켜 구축할 수 있다. 본 연구에서는 특허의 제목과 초록 데이터를 사용하여 fastText 모델을 학습한다. 이를 활용하여 특허 텍스트와 비즈니스 영역의 단어 간의 유사도는 식 (8)과 같이 계산한다(Ko *et al.*, 2020).

$$Sim(PA, BA) = \min_{j}(\max_{i}(sim(w_{PA,i}, w_{BA,j}))) \quad (8)$$

여기서 $w_{PA,i}$는 특허 PA의 제목과 초록의 i번째 단어이며, $w_{BA,j}$는 비즈니스 영역 BA의 j번째 단어이고, $sim(w_{PA,i}, w_{BA,j})$는 두 단어 $w_{PA,i}$와 $w_{BA,j}$ 임베딩 벡터 간의 코사인 유사도이다. 즉, $sim(w_{PA,i}, w_{BA,j})$의 최대값을 계산하여 $w_{BA,j}$와 PA사이의 연관성을 식별하고, $\max_{i}(sim(w_{PA,i}, w_{BA,j}))$의 최솟값을 계산하여 <Figure 5>의 (b)와 같이 PA와 BA간의 링크를 식별한다.



(a) The Architecture of Skip-gram (b) Patent-Trademark Linkage

Figure 5. Linking Patent-Trademark Data based on Embedding

4. 제안 방법론의 실시 예

본 장에서는 글로벌 기술 기반 기업들의 미국 특허와 상표 데이터를 기반으로 상표 지정상품이 구체적으로 유형화되어 있지 않은 경우 이를 비즈니스 영역으로 식별하고, 특허-상표 데이터를 연결함으로써 비즈니스 인텔리전스 소재로 활용하는 일 실시 예를 보인다.

4.1 데이터 수집 및 전처리

본 실험에서는 글로벌 기술 기반 기업들을 선정하여 경쟁자 유형에 따라 잠재적, 외부, 내부 경쟁자 유형으로 분류하였고(Wnat *et al.*, 2011), 한국특허정보원(korea Institute of patent information, 이하 KIPRIS)의 IP 검색 서비스를 활용하여, <Table 5>와 같이 각 기업의 최근 5년(2018-2022) 미국 특허 데이터와 상표 데이터를 확보하였다.

<Figure 6>은 수집된 글로벌 기술 기반 기업들의 특허와 상표 데이터 비율이다. 상표에 비해 월등히 많은 특허의 비율로 인해 y축의 IP 건수를 logarithmic scaling 하였고, 대부분의 기업이 상표에 비해 특허를 더 많이 보유한 것을 확인할 수 있다.

<Table 6>은 수집된 상표 데이터의 Nice Classification(이하 NCL)별 분포이다. 정기적으로 개정되고 있는 NCL은 총 45개 류(class)로 구분되고, 하위의 지정상품을 포함하며, 상품은 1~34류, 서비스업은 35~45류 중 하나에 속한다. 이처럼 기업들은 상표를 등록할 때, 해당 상표를 사용하고자 하는 비즈니스와 연관되어 있는 류구분, 유사군 코드 및 상품/서비스를 설정하게 된다. <Figure 7>은 NCL별 누적 분포를 나타내는데, 그중에서도 9류(전기 및 과학 장치), 42류(컴퓨터, 과학 및 법률), 35류(광고 및 비즈니스 서비스), 41류(교육 및 엔터테인먼트 서비스), 12류(차량) 순으로 비즈니스 영역 Top-5에 해당하는 것을 알 수 있다.

글로벌 기술 기반 기업별로 Top-5 비즈니스 영역의 분포를 살펴보면 <Figure 8>과 같다. 이를 통해서 삼성과 애플은 '전기 및 과학 장치' 비즈니스 영역에서, 아마존과 애플은 '컴퓨터, 과학 및 법률' 분야에서, 또한 아마존과 쿠팡은 '광고 및 비즈니스 서비스' 분야에서, 벤츠와 아우디는 '차량'분야가 주된 사업영역임을 확인할 수 있고, 아마존과 애플은 '교육 및 엔터테인먼트 서비스' 분야로도 사업을 다각화하고 있음을 통계적으로 확인할 수 있다.

이와 같은 글로벌 기술 기반 기업의 기술-비즈니스 영역을 연계를 위해서는 먼저 상표 지정상품의 전처리가 필요하다. <Figure 9>는 수집된 상표 지정상품의 예시로서 특별

Table 5. Summary of IP Database by Competitor Type

Type of Competitor		IP Database	
		Patent	Trademark
Potential	Amazon Technologies, Inc.	5,795	478
	Coupang Corp.	370	60
External	Apple Inc.	14,540	362
	SAMSUNG ELECTRONICS CO., LTD.	19,693	870
Internal	AUDI AG	574	21
	Bayerische Motoren Werke AG	8	2
	Mercedes-Benz Group AG	53	202
Total		41,033	1,995

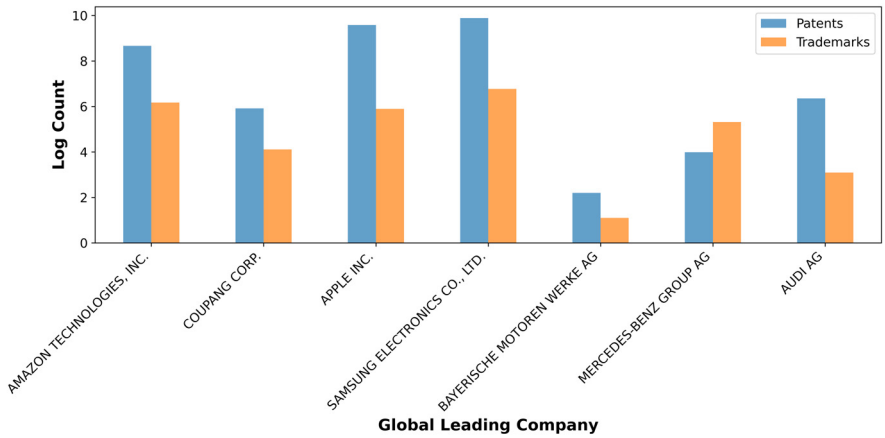


Figure 6. Ratio of Global Technology Based Firms by IP

한 규칙 없이 서술식으로 기재되어 있어 해당 상표로 비즈니스 영역을 식별하기에는 어려움이 있다. 이에 <Figure 10>과 같이 텍스트 전처리 과정을 거쳐 텍스트 분석이 가능하게 한다.

Table 6. Distribution by NCL in Trademark Data

Class	Goods & Services	Distribution
1	Chemical Producs	2
2	Paints & Varnishes	3
3	Cosmetics & Cleaning Substances	23
4	Industrial Oils and Lubricants	5
5	Medicines	11
6	Common Metals & Alloys	11
7	Machine Tools	94
8	Hand Tools	4
9	**Electric and Scientific Devices**	**1,264**
10	Medical Apparatus	24
11	Environmental Control Apparatus	118
12	**Vehicals**	**178**
13	Firearms	1
14	Jewellery	31
15	Musical Instruments	1
16	Stationery and Paper Goods	30
17	Rubber Goods	2
18	Leather Goods	16
19	Building Materials	1
20	Furniture and Materials not otherwise specified	12
21	Houseware and Glass	20
22	Ropes and Fibers	1
23	Yarns and Thread	1
24	Fabriscs	7
25	Clothing and Footware	39
26	Fancy goods usch as Lace and Embroidery	1
27	Carpets and Floor Coverings	5
28	Toys and Sporting Goods	23
29	Meats and Processef Foods	16
30	Staple foods including Flour, cereals, bread etc.	16
31	Natural Agricultural Products	8
32	Light Beverages including Beer	7
33	Wines and Spirits	1
34	Tobacco Products	1
35	**Advertising and Business Services**	**216**
36	Insurance and Financial Services	60
37	Building Contruction & Repair Services	34
38	Telecommunication Services	138
39	Transportation and Storage Services	60
40	Material Treatment Services	24
41	**Education and Entertainment Services**	**214**
42	**Computer, Scientific and Legal**	**349**
43	Restaurants and Food Service	14
44	Medical and Veterinary Services	15
45	Personal and Social Services	48

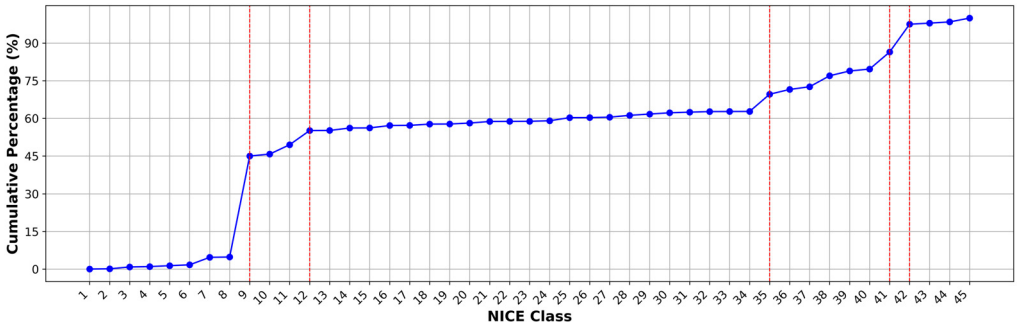


Figure 7. Cumulative Percentage of NCL Distribution

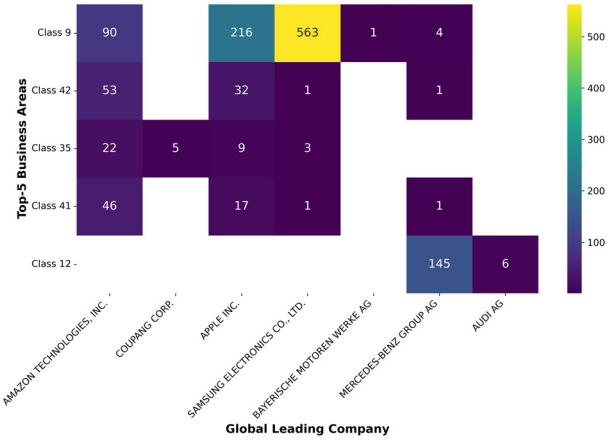


Figure 8. Distribution of Business Area Top-5 by Global Technology Based Firms

'(Based on Use in Commerce) Wrist-mounted smartphones; Smart watches that communicate data to smartphone; Smartphones in the shape of a watch; Wearable digital electronic devices incorporating smartphones in the shape of a watch in the nature of smart watches; Wearable digital electronic devices, comprised primarily of smart watchbands that communicate data to smartphones and comprised primarily of software and display screens for viewing, sending and receiving texts, emails, data and information to smartphones all in the nature of smart watches; Smart watches that communicate data to smartphones, through Internet websites and other computer and electronic communication networks; Smart watches incorporating cameras and MP3 players, and that communicate data to smartphones and PDAs; Wearable computers in the nature of smartwatches; Wearable computer peripherals; Wearable digital electronic communication devices in the nature of smartphones; Computer application software for mobile devices for interworking with mobile device accessories, namely, earphone, headphones, rechargeable batteries, battery chargers, leather cases for mobile phones having NFC chips, flip covers for mobile phones having NFC chips, wireless speakers and 3D eye glasses; Bands for smartphones in the shape of a watch in the nature of smartwatch bands; Smartwatch band straps for smartphones in the shape of a watch; Battery chargers for smartphones in the shape of a watch; Electronic cable for smartphones in the shape of a watch, (Based on 44(e)) Watch bands and straps; Parts and fittings for wristwatches; Watch cases being parts of watches'

Figure 9. Raw Example of an Trademark Goods and Services in the Original Form

Step 1. Punctuation removal and lower case conversion

'based on use in commerce wristmounted smartphones smart watches that communicate data to smartphone smartphones in the shape of a watch wearable digital electronic devices incorporating smartphones in the shape of a watch in the nature of smart watches wearable digital electronic devices comprised primarily of smart watchbands that communicate data to smartphones and comprised primarily of software and display screens for viewing sending and receiving texts emails data and information to smartphones all in the nature of smart watches smart watches that communicate data to smartphones through internet websites and other computer and electronic communication networks smart watches incorporating cameras and mp3 players and that communicate data to smartphones and pdas wearable computers in the nature of smartwatches wearable computer peripherals wearable digital electronic communication devices in the nature of smartphones computer application software for mobile devices for interworking with mobile device accessories namely earphone headphones rechargeable batteries battery chargers leather cases for mobile phones having nfc chips flip covers for mobile phones having nfc chips wireless speakers and 3d eye glasses bands for smartphones in the shape of a watch in the nature of smartwatch bands smartwatch band straps for smartphones in the shape of a watch battery chargers for smartphones in the shape of a watch electronic cable for smartphones in the shape of a watch based on 44e watch bands and straps parts and fittings for wristwatches watch cases being parts of watches'

Step 2. Stopword removal

based use commerce wristmounted smartphones smart watches communicate data smartphone smartphones shape watch wearable digital electronic devices incorporating smartphones shape watch nature smart watches wearable digital electronic devices comprised primarily smart watchbands communicate data smartphones comprised primarily software display screens viewing sending receiving texts emails data information smartphones nature smart watches smart watches communicate data smartphones internet websites computer electronic communication networks smart watches incorporating cameras mp3 players communicate data smartphones pdas wearable computers nature smartwatches wearable computer peripherals wearable digital electronic communication devices nature smartphones computer application software mobile devices interworking mobile device accessories namely earphone headphones rechargeable batteries battery chargers leather cases mobile phones nfc chips flip covers mobile phones nfc chips wireless speakers 3d eye glasses bands smartphones shape watch nature smartwatch bands smartwatch band straps smartphones shape watch battery chargers smartphones shape watch electronic cable smartphones shape watch based 44e watch bands straps parts fittings wristwatches watch cases parts watches

Step 3. Tokenization and lemmatization

'based', 'use', 'commerce', 'wristmounted', 'smartphones', 'smart', 'watch', 'communicate', 'data', 'smartphone', 'smartphones', 'shape', 'watch', 'wearable', 'digital', 'electronic', 'device', 'incorporating', 'smartphones', 'shape', 'watch', 'nature', 'smart', 'watch', 'wearable', 'digital', 'electronic', 'device', 'comprised', 'primarily', 'smart', 'watchband', 'communicate', 'data', 'smartphones', 'comprised', 'primarily', 'software', 'display', 'screen', 'viewing', 'sending', 'receiving', 'text', 'email', 'data', 'information', 'smartphones', 'nature', 'smart', 'watch', 'smart', 'watch', 'communicate', 'data', 'smartphones', 'internet', 'website', 'computer', 'electronic', 'communication', 'network', 'smart', 'watch', 'incorporating', 'camera', 'mp3', 'player', 'communicate', 'data', 'smartphones', 'pda', 'wearable', 'computer', 'nature', 'smartwatches', 'wearable', 'computer', 'peripheral', 'wearable', 'digital', 'electronic', 'communication', 'device', 'nature', 'smartphones', 'computer', 'application', 'software', 'mobile', 'device', 'interworking', 'mobile', 'device', 'accessory', 'namely', 'earphone', 'headphone', 'rechargeable', 'battery', 'battery', 'charger', 'leather', 'case', 'mobile', 'phone', 'nfc', 'chip', 'flip', 'cover', 'mobile', 'phone', 'nfc', 'chip', 'wireless', 'speaker', '3d', 'eye', 'glass', 'band', 'smartphones', 'shape', 'watch', 'nature', 'smartwatch', 'band', 'smartwatch', 'band', 'strap', 'smartphones', 'shape', 'watch', 'battery', 'charger', 'smartphones', 'shape', 'watch', 'electronic', 'cable', 'smartphones', 'shape', 'watch', 'based', '44e', 'watch', 'band', 'strap', 'part', 'fitting', 'wristwatch', 'watch', 'case', 'part', 'watch'

Step 4. Vectorization

[[0.03737175, 0.03737175, 0.03737175, 0.14948702, 0.07474351, 0.11211526, 0.03737175, 0.03737175, 0.07474351, 0.07474351, 0.07474351, 0.03737175, 0.14948702, 0.07474351, 0.07474351, 0.14948702, 0.03737175, 0.18685877, 0.18685877, 0.11211526, 0.03737175, 0.03737175, 0.18685877, 0.03737175, 0.03737175, 0.03737175, 0.03737175, 0.03737175, 0.07474351, 0.03737175, 0.03737175, 0.03737175, 0.03737175, 0.14948702, 0.03737175, 0.03737175, 0.18685877, 0.03737175, 0.07474351, 0.07474351, 0.03737175, 0.03737175, 0.07474351, 0.03737175, 0.07474351, 0.03737175, 0.03737175, 0.03737175, 0.03737175, 0.22423053, 0.22423053, 0.03737175, 0.44846106, 0.07474351, 0.03737175, 0.07474351, 0.03737175, 0.07474351, 0.03737175, 0.03737175, 0.03737175, 0.52320456, 0.03737175, 0.18685877, 0.03737175, 0.03737175, 0.03737175, 0.03737175]]

Figure 10. An Example of Pre-processing for Trademark Goods and Services

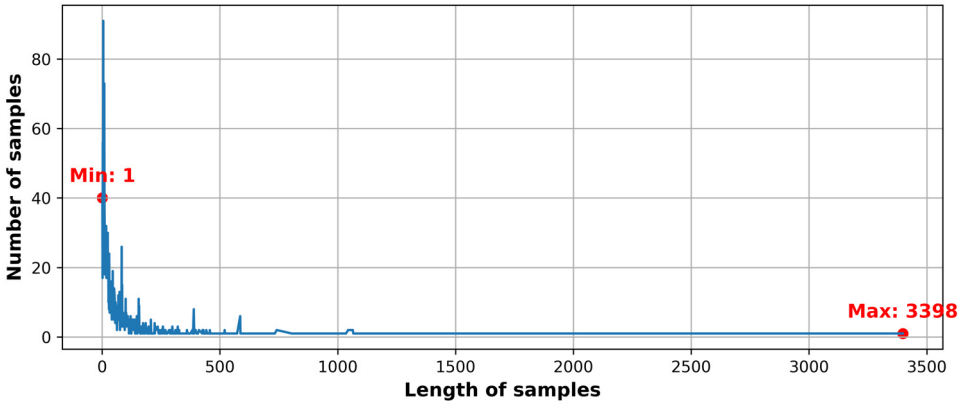


Figure 11. Example of Preprocessed Trademark Data

Table 7. Summary of Preprocessed Trademark Data

Criteria	Word Count	Trademark Index	Right Holder	NICE Class	Trademark Goods & Services
Min	1	87	Amazon Technologies, Inc.	18	['umbrella']
Max	3,398	501	Coupang Corp.	9, 35, 36, 38	['computer', 'retail', 'mobile', 'board', 'software', 'service', 'wireless', 'information', 'featuring', 'work', ⋯]
Mean	133	otherwise			

전처리된 상표 지정상품에 대한 정보를 요약하면 <Figure 11>과 <Table 7>과 같다. 수집된 상표의 지정상품을 구성하는 단어는 평균 133개이지만, 최대 3,398개까지 다양하다. 즉, 이러한 단어들을 유형화하여 유사군으로 표상하는 작업이 필요하다.

4.2 기술 기반 기업의 상표 지정상품 유사군 표상

본 실험에서는 LDA 기반의 토픽 모델링을 위해 Python의 Gensim 라이브러리에서 제공하는 'LdaMulticore' 클래스를 수정 활용한다. 추론 과정에서 디리클레 하이퍼파라미터인 α는 상표 집합(corpus)에 의해 정해진 기본값을 사용하였으며, β또한 기본값인 0.1을 사용한다. 또한 내장된 깁스 샘플링(gibbs samping)을 이용하여 기본값인 100회 반복(iterations)한다. 토픽의 수(z_k)를 결정하기 위해서 상표 지정상품이 NCL 45개 류(class)에 대한 내용을 담고 있는 점을 고려하여 1부터 45까지 coherence score를 평가한다. 평가 결과 <Figure 12>와 같이 일관성 점수가 가장 높은 6을 토픽의 수로 설정한다. 그러나 정량지표 외에도 토픽의 해석이 중요한 특정 분야에 관한 연구에서는 도메인 전문가들의 판단을 고려한 정성적 평가도 필요할 것으로 사료된다(Lee and Kang, 2018). 상기 학습 과정을 통해 LDA 기반으로 추론된 토픽이 지칭하는 단어 즉, 지정상품은 <Table 8>과 같다. 이때 토픽 별 단어의 수(w_i)는 기본값인 10으로 설정하였으나, 토픽의 수와 마찬가지로 분석하고자 하

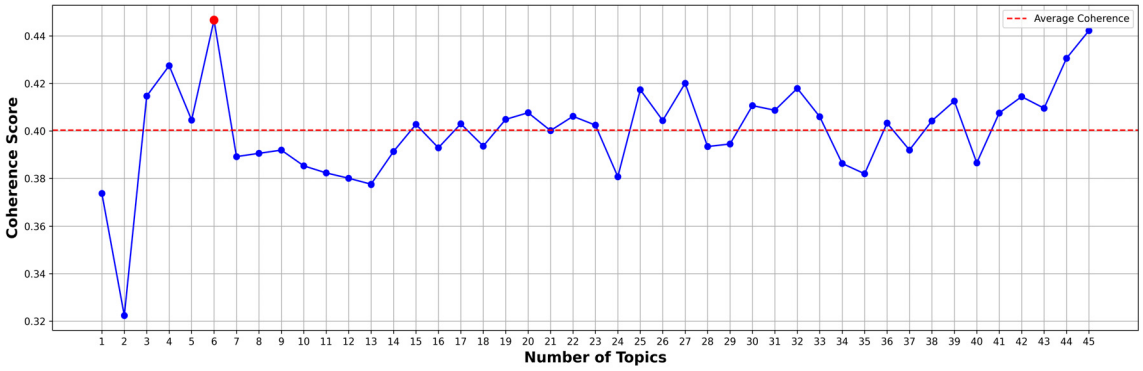


Figure 12. LDA Coherence Results Over Different Number of Topics

Table 8. Topic Modeling Results of Best Model

Topic	1^{st} Word(probability)	⋯	4^{st} Word(probability)	⋯	8^{st} Word(probability)	⋯	10^{st} Word(probability)
1	television(0.068)		display(0.024)		digital(0.013)		software(0.012)
2	smartphones(0.027)		mobile(0.016)		wearable(0.014)		battery(0.013)
3	software(0.018)	⋯	providing(0.009)	⋯	application(0.007)	⋯	game(0.007)
4	vehicle(0.042)		thereof(0.030)		audio(0.009)		electric(0.008)
5	shirt(0.009)		baby(0.006)		short(0.005)		jacket(0.005)
6	cleaner(0.022)		memory(0.015)		household(0.013)		washing(0.013)

는 분야나 도메인 전문가의 판단에 따라 조정(adjustment)될 수 있다. 수집한 상표 데이터 1,995건의 각 토픽별 분포는 <Figure 13>과 같이 토픽 3에 가장 많은 분포를 보이고 있는데, 토픽 3이 지칭하는 단어는 대부분 전기가 필요한 지정상품들이고, 앞서 <Table 6>과 <Figure 7>을 통해 살펴본 것과 같이 수집된 대부분의 상표 데이터가 NCL Class 9 (‘전기 및 과학장치’)에 속하는 것과 같은 결과이다.

이처럼 LDA 토픽 모델링을 활용하여 상표 지정상품을 유형화함으로써 유사군을 표상할 수 있다. 그러나 <Table 7>의 501번째 상표와 같이 하나의 상표가 다수 개의 NCL로 구성되고, 지정상품이 수 천 가지에 이르는 경우, 해당 상표는 다수 개의 유사군으로 표상되는 것이 바람직하다. 이를 위해서 본 실험에서는 <Table 9>와 같이 상표(d_i)가 각 주제(z_k)별로 속할 확률이 주제별 확률값의 평균 이상인 경우(δ) 해당 지정상품의 유사군으로 선택한다. 이때 임곗값 δ는 데이터의 도메인이나, 전문가의 의견에 따라 조정할 수 있다.

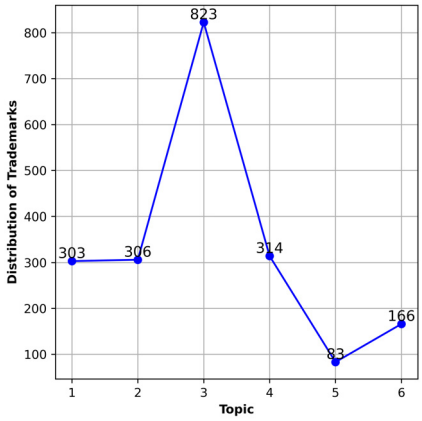


Figure 13. Distribution of Trademarks by Topics

Table 9. Topic Distribution by Item

Trademark Index	NICE Class	Topic			
		Rank	Score	Threshold (δ)	Selection
180	2, 3, 5, 6, 16, 21, 29, 30, 31, 32, 35	5	0.8450	0.3311	5
		6	0.1231		
		3	0.0252		
476	9, 35, 36, 39, 41, 42, 43, 44, 45	3	0.9881	0.4996	3
		4	0.0111		
494	1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 37, 38, 40, 41, 43, 44	3	0.3379	0.1667	3, 5, 6
		5	0.2967		
		6	0.2336		
		4	0.0766		
		2	0.0294		
		1	0.0258		
499	38, 39	3	0.7735	0.4996	3
		6	0.2257		
661	9	3	0.4486	0.3318	3
		4	0.3012		
		2	0.2457		
923	10, 35, 36, 39, 41, 42, 43, 44, 45	3	0.8890	0.3325	3
		2	0.0600		
		6	0.0484		

Table 9. Topic Distribution by Item(Continued)

Trademark Index	NICE Class	Topic			
		Rank	Score	Threshold (δ)	Selection
1098	9, 14	2	0.9282	0.4977	2
		3	0.0673		
1802	3, 6, 7, 8, 9, 11, 14, 16, 18, 20, 21, 24, 25, 27, 28	3	0.4490	0.1667	3
		4	0.1545		
		6	0.1537		
		5	0.1461		
		2	0.0488		
		1	0.0479		
1808	3, 6, 9, 11, 12, 14, 16, 18, 20, 21, 24, 25, 27, 28, 35, 36, 37, 39, 041, 43	3	0.3929	0.1667	3, 4, 6, 5
		4	0.2079		
		6	0.1745		
		5	0.1736		
		2	0.0410		
		1	0.0101		
1962	12	4	0.8611	0.1667	4
		2	0.0278		
		6	0.0278		
		3	0.0278		
		5	0.0278		
		1	0.0278		

예를 들어서, <Table 9>의 494번째 상표는 6가지 토픽에 모두 속할 확률을 가지며, 각 주제별 평균값인 0.1667 이상인 토픽 3, 5, 6에 대해서만 유사군으로 본다. 아이템별 유사군 표상 결과를 이분 네트워크(bipartite network)로 시각화하면 <Figure 14>와 같다.

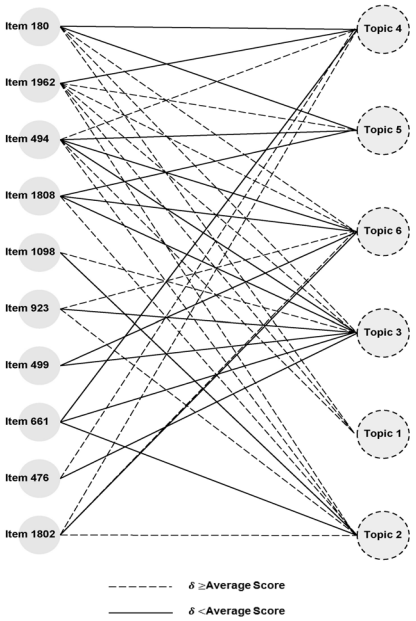


Figure 14. Similar Group Representation Results by Item

4.3 기술 기반 기업의 비즈니스 영역 식별

본 절에서는 표상된 유사군과 SIC의 유사도에 기반하여 비즈니스 영역을 식별한다. 식별에 앞서 사전에 수집된 1,005건의 SIC 데이터를 통해 산업별 분포를 분석한 결과는 <Figure 15>와 같다. 전체 산업은 10개 부문으로 나뉘며 각 부문에는 83개의 주요 집단, 416개 산업집단과 1,005개의 SIC 코드를 갖는다. 그러나 본 연구에서 활용되는 SIC 레이블은 1,003개로 <Table 10>과 같이 산업 부문 D와 I그룹에 중복된 레이블이 존재하기 때문으로 확인되었다. SIC 레이블은 <Figure 16>과 같이 최대 10개의 단어로 구성되는데 이러한 SIC 레이블과 표상된 유사군의 단어 임베딩을 통해 유사도를 계산함으로써 비즈니스 영역을 식별한다.

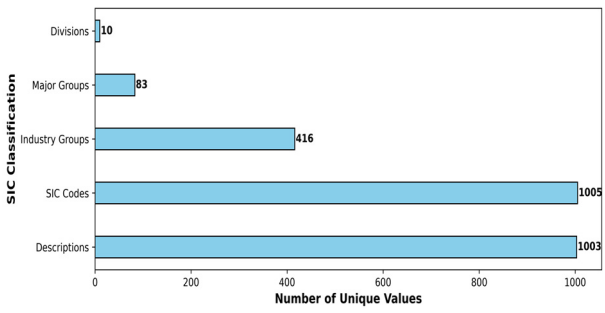


Figure 15. Distribution by Industry

Table 10. Duplicate SIC Description by Industry

Industry Title	Division	Major Group	Industry Group	SIC	Description
Manufacturing	D	38	385	3851	Ophthalmic Goods
			386	3861	Photographic Equipment and Supplies
Service	I	50	504	5043	Photographic Equipment and Supplies
				5048	Ophthalmic Goods

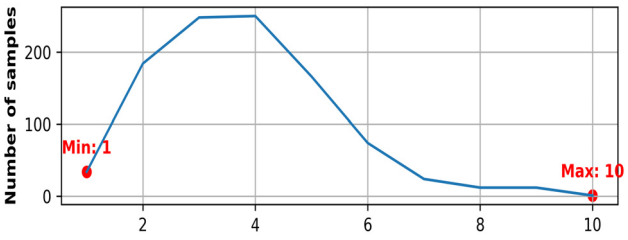


Figure 16. Example of Preprocessed SIC Description

SIC 레이블도 상표 지정상품과 동일한 방식으로 텍스트 전처리를 수행한 후 이를 기반으로 fastText 모델을 학습한다. 임베딩된 유사군 벡터와 SIC 벡터의 유사도를 비교한 결과는 <Table 11>과 <Table 12>와 같다. <Table 11>에서 토픽 6은 SIC 5999와 0.5947의 유사도로 매칭됨을 알 수 있다. <Table 12>는 매칭 과정에서 가중된 유사도를 계산한 결과로 유사군과 SIC 간의 최대 유사도와 최소 유사도가 향상된 것을 정량적으로 확인할 수 있다. 또한 주황색으로 음영 처리된 부분을 구

Table 11. Unweighted Similarity between Topic and SIC with fastText

Topic	Unweighted Similarity				SIC			
	Min	Max	Avg.	Std.	Division	Major Group	Industry Group	SIC
1	-0.2970	0.5426	0.0988	0.1153	F	50	506	5064
2	-0.2281	0.3517	0.0816	0.1007	B	12	1234	1241
3	-0.2040	0.5864	0.2074	0.1475	I	87	874	8741
4	-0.1971	0.5787	0.1261	0.1260	F	50	501	5015
5	-0.2528	0.5258	0.0759	0.1071	D	23	232	2321
6	-0.2230	0.5947	0.1796	0.1242	G	59	599	5999

Table 12. Weighted Similarity between Topic and SIC with fastText

Topic	Weighted Similarity				SIC			
	Min	Max	Avg.	Std.	Division	Major Group	Industry Group	SIC
1	-0.1986	0.5791	0.1337	0.1171	F	50	506	5064
2	-0.2175	0.3548	0.0714	0.0995	B	12	1234	1241
3	-0.1804	0.6304	0.1907	0.1469	I	87	874	8741
4	-0.2194	0.6354	0.1234	0.1266	F	50	501	5015
5	-0.2709	0.5691	0.0813	0.1102	D	23	232	2321
6	-0.1972	0.6031	0.1799	0.1214	D	36	363	3635

Table 13. Mapping Results of Business Areas based on Weighting

Topic			Unweighted			Weighted		
Index	Word	Prob	SIC	Description	Similarity	SIC	Description	Similarity
6	cleaner	0.022	5999	Miscellaneous Retail Stores, Not Elsewhere Classified	0.5947	3635	Household Vacuum Cleaners	0.6031
	vacuum	0.021						
	drive	0.016						
	memory	0.015						
	blank	0.014						
	retail	0.014						
	store	0.013						
	household	0.013						
	flash	0.013						
	washing	0.013						

체적으로 정리한 <Table 13>을 보면 SIC 5999의 경우 토픽 6에 포함된 단어는 'Retail'과 'Store' 두 가지 뿐이고, SIC 3635는 레이블의 세 가지 단어 모두 토픽 6에 포함된다. 그중에서도 'Cleaner'와 'Vacuum' 두 단어는 토픽 6에서 높은 가중치를 지님으로 매칭 결과가 신뢰할 수 있음을 확인한다. 이처럼 표상된 유사군과 SIC 레이블의 가중된 코사인 유사도 기반의 매칭은 정교한 비즈니스 영역의 식별을 가능케 한다.

상기 실험 결과를 토대로 <Table 9>의 아이템별 비즈니스 영역을 식별한 결과는 <Table 14>와 같고 <Figure 16>은 이를 삼분 네트워크(tripartite network)로 시각화한 결과이다. 이를 통해서 상표 지정상품이 구체적으로 지정되지 않은 경우에도 해당 상표의 지정상품이 어떠한 제품, 서비스 영역에서 활용하고자 하는지 식별할 수 있다.

4.4 기술 기반 기업의 기술-비즈니스 영역 연결

앞서 상표 지정상품 텍스트를 분석하여 유사군을 표상하고 이를 SIC와 매칭함으로써 비즈니스 영역을 식별하였다. 본 절에서는 특허-상표 데이터 연결 기반의 비즈니스 인텔리전스 초기 연구(Ko *et al.*, 2020)를 확장한다는 점에서 기존 연구와 동일한 방식으로 글로벌 기술 기반 기업의 기술-비즈니스 영역 연계를 수행하였고 그 결과는 <Table 15>와 같다. 예를 들어, 상표명이 'COUPANG'인 경우 유사군 코드를 통해 실제로 어떠한 제품이나 서비스에 해당하는지 유추해 볼 수 있다. 그러나 유사군이 구체적으로 지정되어 있지 않은 경우, 제안 방법론을 통해 매칭된 SIC를 통해 'Management Service'가 주된 비즈니스 영역임을 식별할 수 있다. 또한 이를 통해 쿠팡이 보유

Table 14. Identify Business Areas by Item

Trademark Index	NICE Class	Business Area		
		Topic	SIC	Description
180	2, 3, 5, 6, 16, 21, 29, 30, 31, 32, 35	5	2321	Men's and Boys' Shirts, Except Work Shirts
476	9, 35, 36, 39, 41, 42, 43, 44, 45	3	8741	Management Services
494	1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 37, 38, 40, 41, 43, 44	3	8741	Management Services
		5	2231	Men's and Boys' Shirts, Except Work Shirts
		6	3635	Management Services
499	38, 39	3	8741	Management Services
661	9	3	8741	Management Services
923	10, 35, 36, 39, 41, 42, 43, 44, 45	3	8741	Management Services
1098	9, 14	2	1241	Coal Mining Services
1802	3, 6, 7, 8, 9, 11, 14, 16, 18, 20, 21, 24, 25, 27, 28	3	8741	Management Services
1808	3, 6, 9, 11, 12, 14, 16, 18, 20, 21, 24, 25, 27, 28, 35, 36, 37, 39, 041, 43	3	8741	Management Services
		4	5015	Motor Vehicle Parts, Used
		6	3635	Household Vacuum Cleaners
		5	2321	Household Vacuum Cleaners
1962	12	4	5015	Motor Vehicle Parts, Used

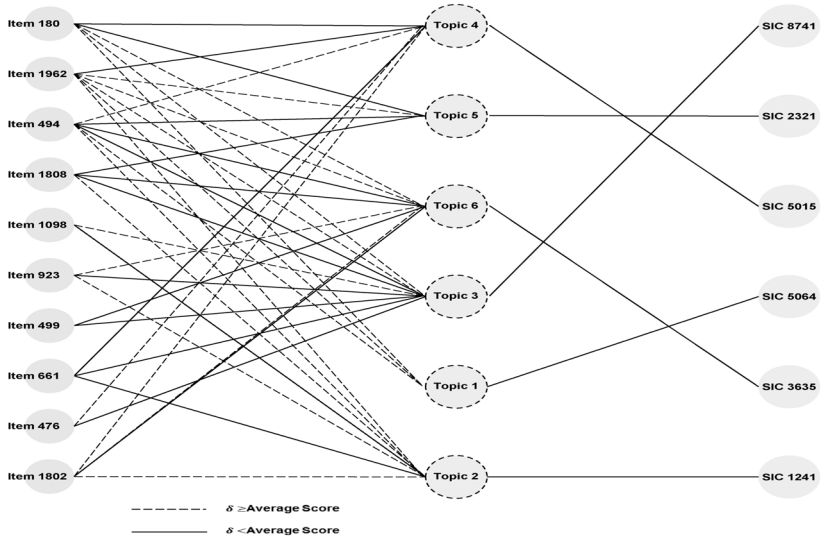


Figure 17. Business Area Identification Results by Item

Table 15. Examples of Linked Business Areas by Patent

Application Number Title	Source (sim)	Trademark name	Business Area		
			Division	SIC	Description
16558570 Providing replacement items for discontinued items (Adrian Bell, Andre Wyatt, 2019)	Title (0.8945) Abstract (0.8117)	ALEXA	Service	8741	Management Services
16380239 Systems and methods for machine-learning assisted inventory placement (Xin Shi, 2019)	Title (0.7571) Abstract (0.8408)	COUPANG	Service	8741	Management Services
			Manufacturing	2231	Men's and Boys' Shirts, Except Work Shirts
			Manufacturing	3635	Management Services
17910116 METHOD FOR ADAPTING A FUNCTIONALITY OF A VEHICLE THAT IS AUTOMATICALLY CARRIED OUT BY AN ASSISTANCE SYSTEM, AND A DRIVER ASSISTANCE SYSTEM (PETERS *et al.*, 2021)	Title (0.7824) Abstract (0.7112)	S 560	Wholesale Trade	5015	Motor Vehicle Parts, Used

한 특허 중 'Systems and methods for machine-learning assisted inventory placement'가 해당 상표와 연관된 기술임을 유추할 수 있다. 이로써 기술 기반 기업이 실제 비즈니스 환경에서의 어떤 제품이나 서비스와 경쟁하는지 파악하고, 해당 기업이 비즈니스 영역에서 활용하는 기술과 활용하고 있지 않은 기술을 필터링 함으로써 비즈니스 인텔리전스를 제공할 수 있다.

5. 결론 및 향후 연구

속지주의를 따르는 상표의 특성상 국가별로 상표 제도에 따른 차이로 기존의 방법으로는 특허-상표를 연계한 비즈니스 인텔리전스를 수행하는 데 한계가 있다. 본 연구에서는 상표 데이터 지정상품의 유사군이 구체적으로 지정되어 있지 않은 경우, 특허-상표를 연계를 위해 상표 지정상품의 텍스트를 분석하여 비즈니스 영역을 식별하는 방법을 제시한다. 이를 위해서 LDA 기반의 토픽 모델링을 이용하여 상표 지정상품의 유사군을 표상하고, 다음으로 표상된 유사군과 SIC의 유사도를 기반으로 비즈니스 영역을 식별한다. 이때 표상된 유사군을 지칭하는 단어별 가중치를 고려한 가중된 코사인 유사도를 사용하여 정교한 비즈니스 영역이 매칭되도록 한다. 이를 통해 글로벌 기술 기반 기업의 특허 데이터와 상표 데이터를 연계하여 비즈니스 인텔리전스 소재로 활용 가능하게 한다. 실험결과 표상된 유사군은 수집된 상표 데이터의 NCL Class 정보와 대부분 일치하는 점으로 보아 일관성 있게 추론되었음을 확인한다. 또한 비즈니스 영역을 식별하는 과정에서 fastText 모델로 임베딩된 벡터에 표상된 유사군의 단어별 가중치를 반영하여 SIC와 매칭 유사도가 향상됨을 확인한다. 이를 통해 글로벌 기술 기반 기업의 특허-상표 데이터의 연계가 가능하고, 실제 비즈니스 환경을 고려한 제품이나 서비스 수준에서의 비즈니스 분석이 가능케 된다.

그러나 제안 방법론에서 활용된 모델의 특성상 분석대상기업이 바뀌거나 신규특허가 추가되는 등 실시간 분석에는 한계가 있다. 먼저 유사군을 표상하는 과정에서 LDA 모델의 특성상 최적의 토픽 개수를 사전에 정해야 하며, 비 계층 구조로 구성되어 있어 토픽 간의 연관성을 고려할 수 없는 한계가 있다. 또한 디리클레 하이퍼파라미터는 유사군을 지칭하는 지정상품의 구성에 영향을 미친다. 따라서 향후 연구에서는 이러한 한계를 보완할 수 있는 토픽 모델이 고려되어야 할 것이다. 둘째로, 비즈니스 영역을 식별하는 과정에서 사용된 임베딩 모델이 static하여 동일한 형태의 단어일지라도 문장의 맥락에 따라 변화하는 단어들의 의미를 반영할 수 없다는 한계가 있다(Huang *et al.*, 2012). 따라서 추후 연구에서는 contextual 기반의 임베딩 모델을 고려할 여지가 있다. 향후 연구에서는 상기 한계점을 보완하여 급변하는 경쟁 환경에서 실시간으로 활용 가능한 방법이 고려되어야 하겠다. 더불어 상표 데이터 외에도 다양한 외부 데이터를 연계하여 비즈니스 분석에 활용한다면 더욱 고도화된 비즈니스 인텔리전스를 이룩할 것이라고 사료된다.

참고문헌

Barroso, A., Giarratana, M. S., and Pasquini, M. (2019), Product portfolio performance in new foreign markets: The EU trademark dual system, *Research Policy*, **48**(1), 11-21.

Blei, D. M. and Lafferty, J. D. (2009), Topic models. In Text mining (pp. 101-124). Chapman and Hall/CRC.

Chen, H., Chiang, R. H., and Storey, V. C. (2012), Business intelligence and analytics: From big data to big impact, *MIS Quarterly*, 1165-1188.

Choi, J. and Yoon, J. (2021), Detecting Promising Business Areas through Trademark Designated Goods Text Analysis, *Proceedings of the Korean Society for Management and Science*, 18-36.

Choi, J., Jeong, B., and Yoon, J. (2019), Technology opportunity discovery under the dynamic change of focus technology fields: Application of sequential pattern mining to patent classifications, *Technological Forecasting and Social Change*, **148**, 119737.

Chudnovsky, D. (1983), Patents and trademarks in pharmaceuticals, *World Development*, **11**(3), 187-193.

Davenport, T. H. (2006), Competing on analytics, *Harvard Business Review*, **84**(1), 98.

Ferreira, V. and Godinho, M. (2011), Building an innovation function with patents and trademarks: Evidence from Portuguese regional innovation systems, *Innovation, Strategy, and Structure-Organizations, Institutions, Systems and Regions*, 1-29.

Hao, Y. S., Shih, H. Y., Huang, H. C., and Lin, L. L. (2010, July), Co-opetition of cooperative and competitive relationship: A network analysis approach, In *Picmet 2010 Technology Management For Global Economic Growth*, IEEE, 1-8.

Hofmann, T. (1999, August), Probabilistic latent semantic indexing, In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 50-57.

Gilad, B. and Herring, J. P. (Eds.) (1996), *The art and science of business intelligence analysis*, JAI Press.

Jeong, B. and Yoon, J. (2017), Competitive intelligence analysis of augmented reality technology using patent information, *Sustainability*, **9**(4), 497.

Jeong, H. (2015), A Study on Ontology and Topic Modeling-based Multi-dimensional Knowledge Map Services, *Journal of Intelligence and Information Systems*, **21**(4), 79-92.

Jung, T. H. (2010), A Comparative Study on Japanese Cancellation Trial System of Trademark Registration Based on Non-Use and Its Characteristics, *Journal of Law Research*, **26**(4), 177-199.

Kim, G., Lee, J., Jang, D., and Park, S. (2016), Technology clusters exploration for patent portfolio through patent abstract analysis, *Sustainability*, **8**(12), 1252.

Kim, S. J., Kim, G. W., and Lee, D. H. (2017), A Topic Related Word Extraction Method Using Deep Learning Based News Analysis, In *Proceedings of the Korea Information Processing Society Conference*, Korea Information Processing Society, 873-876.

Ko, N. (2022), A Patent-Trademark Data Linking Approach to Business Analytics for Technology-Based Firms, *Doctoral Dissertation*, Graduate School, Konkuk University.

Ko, N., Jeong, B., Yoon, J., and Son, C. (2020), Patent-trademark linking framework for business competition analysis, *Computers in Industry*, **122**, 103242

Lee, H. J., and Yang, H. (2015, December), Potential competitor identification and competitor analysis by monitoring patent information, In *ISPIM Innovation Symposium*, The International Society for Professional Innovation Management (ISPIM), 1.

Luo, Y. C. and Su, H. N. (2015, August), Investigating technological evolution of mobile telecommunication industry by integrating dynamic competitive analysis and patent analysis, In *2015 Portland International Conference on Management of Engineering and Technology (PICMET)*, IEEE, 2103-2112.

Lu, Y., Mei, Q., and Zhai, C. (2011), Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA, *Information Retrieval*, **14**, 178-203.

Moon, S. (2015), Applicable Law and the liability for the International Indirect Infringement of Trademark, *Korea Association For Info-Media Law*, **19**(3), 111-131.

Oltra, V. and Saint Jean, M. (2009), Variety of technological trajectories in low emission vehicles (LEVs): A patent data analysis, *Journal of Cleaner Production*, **17**(2), 201-213.

Park, M. Choi, H., and Lee, H. (2021), Study on Preprocessing Method Suitable for Korean Aspect Extraction based on Unsupervised Learning: For Childcare Products Reviews, *Journal of the Korean Institute of Industrial Engineers*, **47**(1), 56-67.

Park, Y. and Yoon, J. (2017), Application technology opportunity discovery from technology portfolios: Use of patent classification and collaborative filtering, *Technological Forecasting and Social Change*, **118**, 170-183.

Pargaonkar, Y. R. (2016), Leveraging patent landscape analysis and IP competitive intelligence for competitive advantage, *World Patent Information*, **45**, 10-20.

Salton, G. and McGill, M. J. (1986), *Introduction to modern information retrieval mcgraw hill book company*, New York.

Sandner, P. (2008, June), The Market Value of R&D, Patents, and Trademarks, In *exposé présenté dans le cadre de l'INNO-tec Conference on the Analysis of Trademarks and Brands*, Alicante.

Seip, M., Castaldi, C., Flikkema, M., and De Man, A. P. (2018), The timing of trademark application in innovation processes, *Technovation*, **72**, 34-45.

Shibata, N., Kajikawa, Y., Takeda, Y., and Matsushima, K. (2008), Detecting emerging research fronts based on topological measures in citation networks of scientific publications, *Technovation*, **28**(11), 758-775.

Song, C. H., Han, J. W., Jeong, B., and Yoon, J. (2017), Mapping the patent landscape in the field of personalized medicine, *Journal of Pharmaceutical Innovation*, **12**, 238-248.

Van der Maaten, L. and Hinton, G. (2008), Visualizing data using t-SNE, *Journal of Machine Learning Research*, **9**(11).

Wang, B., Liu, Y., Zhou, Y., and Wen, Z. (2018), Emerging nanogenerator technology in China: A review and forecast using integrating bibliometrics, patent analysis and technology roadmapping methods, *Nano Energy*, **46**, 322-330.

Yoon, J., Kim, M., Kim, D., Kim, J., and Park, H. (2015), Monitoring the change of technological impacts of technology sectors using patent information: the case of Korea, *Industrial Engineering and Management Systems: An Offcial Journal of APIEMS*, **14**(01), 1-15.

Yoon, J. H., Ko, N. U., Jeong, B. K., and Choi, J. W. (2019), Trademark Data-Based Business Intelligence: Applications and Future Research Issues, *Journal of the Korean Institute of Industrial Engineers*, 270-283.

저자소개

윤주호 : 명지전문대학 산업시스템경영공학과에서 학사학위를 취득하고, 한양대학교에서 산업경영공학과 석박사통합과정에 재학 중이다. 주요 관심분야는 통계 학습, 기계 학습, 데이터마이닝, 기술경영이다.

김병훈 : 럿거스 대학교에서 산업시스템공학 박사학위를 취득하고, 현재 한양대학교 산업경영공학과에 재직 중이다. 주요 관심분야는 통계 데이터 마이닝 방법론 개발, 반도체 제조 공정용 데이터 마이닝 모델 개발, 그래프 마이닝이다.