「정보시스템연구」 제30권 제3호
한국정보시스템학회
2021년 9월, pp. 201~216

http://dx.doi.org/10.5859/KAIS.2021.30.3.201

섬유소재 분야 특허 기술 동향 분석: DETM & STM 텍스트마이닝 방법론 활용*

이현상** · 조보근*** · 오세환**** · 하성호*****

〈목 차〉

Ⅰ. 서론
Ⅱ. 관련 연구
Ⅲ. 제안 기법
3.1 DETM
3.2 STM
Ⅳ. 분석
4.1 데이터
4.2 모델링
Ⅴ. 연구 결과
5.1 토픽 분석 결과
5.2 키워드 분석 결과
5.3 메타데이터 분석 결과
Ⅵ. 결론
참고문헌
<Abstract>

Ⅰ. 서 론

최근 전세계에서 생산되는 데이터의 양이 크게 늘어나면서 빅데이터 플랫폼에 대한 관심이 증가하고 있다. 빅데이터 플랫폼이란 데이터를 수집, 저장, 처리, 관리 및 분석 등의 기능을 지원하여 가치있는 새로운 정보를 얻을 수 있는 통합 빅데이터 프로세스를 의미한다(ETRI, 2019). 전세계 빅데이터 플랫폼 시장의 경우 2019년 약 76조원에서 2025년 약 118조원을 전망하고, 국내 시장에서는 2019년 약 1.5조원에서 2025년 약 2.4조원을 전망한다(ETRI, 2019). 국내에서는 섬유소재 분야에서 빅데이터 플랫폼을 구축하기 위해 다이텍연구원에서 2020년부터 사업을 진행 중이다(매일신문, 2020). 다이텍연구원의 소재산업 빅데이터 플랫폼 구축 사업은 2019년 12월부터 시작된 COVID-19로 인해 비대면 섬유 거래 플랫폼을 활성화하고자 한다. 해외에서는 이미 자라, H&M, 나이키, 아마존 등 유명 브랜드에서 섬유소재 관련 빅데

* 이 논문은 2021년도 산업통상자원부 산업혁신기반구축사업 재원으로 수행된 연구임.(P114000015)
** 경북대학교 경영학부 박사과정, coolwin200@gmail.com(주저자)
*** 주식회사 빅웨이브에이아이 이사, iambgjo@gmail.com
**** 경북대학교 경영학부 부교수, sehwano@knu.ac.kr
***** 경북대학교 경영학부 교수, hsh@knu.ac.kr(교신저자)

이터 활용 기술을 적극적으로 도입하고 있다. 기존 연구에서는 빅데이터 분석을 활용하여 특정 분야에 대한 기술 동향을 분석하기도 한다(Amado et al., 2018; Yang et al., 2018; 홍태호 외., 2018; 최병설 외., 2019; 김민수 외., 2020). 따라서 본 연구의 목적은 섬유소재 분야의 특허 텍스트 데이터를 활용하여 해당 기술의 동향을 분석하는 것이다. 특허는 기술 발명자의 권리를 보장하기 위해 해당 기술을 문서화한 것으로, 산업적 유용성을 중요하게 고려한다. Tseng et al. (2007)은 특허 분석을 통한 자동 분석 도구의 필요성을 제시했다. 기존 연구에서도 해당 기술의 동향을 분석하기 위해 특허 문서의 텍스트 데이터를 활용한다(Yoon et al., 2004; Tseng et al., 2007; Noh et al., 2015).

기존 텍스트마이닝 기반의 기술 동향 분석은 논문, 특허, 웹사이트 게시물 등의 데이터를 활용하여 네트워크 분석(network analysis), 토픽 모델링(topic modeling) 기법을 활용한다(Amado et al., 2018; Yang et al., 2018; 홍태호 외., 2018; 최병설 외., 2019; 김민수 외., 2020). 네트워크 분석의 경우 거시적인 관점에서 텍스트를 분석할 수 있지만, 세부적인 키워드를 분석하는 것이 어렵다는 한계점이 있다. 토픽 모델링 분석은 세부 키워드까지 분석하여 문서의 숨겨진 주제를 파악할 수 있는 기법이지만 주제 분류에서 텍스트 데이터의 메타 정보를 활용할 수 없고 연구자가 원하는 키워드를 분석할 수 없다는 한계점이 존재한다. 본 연구는 DETM(Dynamic Embedded Topic Modeling)과 STM(Structural Topic Modeling)을 연계하는 방법론을 통해서 기존 기술 동향 연구의 한계점을 보완하고 심층적으로 섬유소재 분야의 기술 동향을 분석하고자 한다. 본 연구를 통해서 기술 동향 분석 및 섬유소재 분야 관련 연구자들과 실무자들에게 참고점을 제시하고 해당 산업의 기술적 발전 및 빅데이터 플랫폼의 활성화를 기대한다.

Ⅱ. 관련 연구

텍스트마이닝 기술 동향 분석이란 네트워크 분석 및 토픽 모델링과 같은 방법론을 통해 대량의 텍스트 데이터에서 키워드 및 주제를 파악하고 이에 대한 추세를 분석하는 연구를 지칭한다. 여기서 네트워크 분석(network analysis) 기법이란 네트워크와 그래프 이론을 기반으로 노드(node)와 링크(link)를 구성하여 데이터로부터 사회적 구조를 분석하는 것을 말한다(Scott, 1988). 네트워크 분석은 텍스트마이닝 분야에도 이와 같이 적용되어 단어들을 중심으로 각각의 관계를 분석한다. 토픽 모델링(Topic modeling) 분석 기법은 문서의 단어 집합에서 숨겨진 주제를 발견할 수 있는 알고리즘으로 LSA(Latent Semantic Analysis) 연구로 시작되었다(Hofmann, 1999). 최근에는 LDA (Latent Dirichlet Allocation) 방법론을 기반으로 STM (Structural Topic Model) 및 DETM (Dynamic Embedded Topic Model) 등의 응용 기법으로 메타데이터 및 키워드 트렌드를 분석할 수 있는 형태로 발전하고 있다(Blei et al., 2003; Roberts et al., 2013; Dieng et al., 2019). 네트워크 분석과 토픽 모델링 기법은 데이터에 대한 레이블링(labeling) 처리 없이 수행될 수 있고 다양한 시각화 기능을 지원하기 때문에 대량의 텍스트

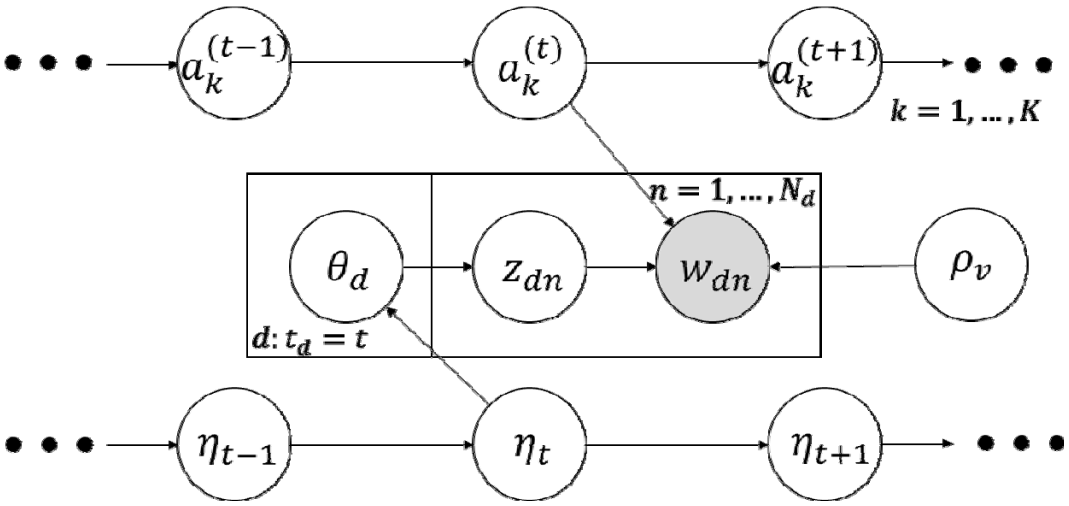
데이터를 분석할 때 유용하게 활용되고 있다.

Kho et al. (2013)은 키워드 기반 네트워크 분석 기법을 통해서 기술경영 분야의 연구 동향을 분석했다. 이 논문에서는 2002년부터 2012년까지 관련 주요 저널 논문의 전문 텍스트 데이터를 수집했다. 네트워크 분석 결과 중심성 분석을 통해 기술 혁신, R&D, 특허, 예측, 기술 이전, 기술, 중소기업 등의 키워드가 근접중심성, 연결중심성, 매개중심성의 수치가 높게 측정되었다. Amado et al. (2018)은 토픽 모델링 기법을 기반으로 마케팅 분야에서의 빅데이터 활용 연구의 관련 동향을 분석했다. 이 연구에서는 빅데이터, 마케팅, 지리적 특성, 제품과 제품군에 대한 차원에서 주제를 추출했다. 2010년부터 2015년까지 게재된 총 1560개의 논문을 분석했으며 결과적으로 실무 영역과 연구 영역이 크게 분류된다고 분석했다. 추가적으로, 마케팅 분야에서의 빅데이터 활용이 아직 초기 연구 단계에 있기 때문에 실제 사업에 대한 직접적인 연구가 필요하다고 주장했다. Kim et al. (2018)은 텍스트마이닝 분석을 통해서 의료정보학 분야의 주제 및 시계열적인 트렌드를 분석했다. 이 논문에서는 2002년부터 2013년까지 의료정보학 저널에 게재된 26,307건의 PubMed 아카이브 논문의 초록 데이터를 수집했다. 다음 단계로 초록 데이터의 말뭉치(corpus)를 분석하여 주제 키워드를 분석했다. 결과적으로 생물의학(biomedical)과 같은 일부 주제는 점유율이 감소하는 반면에, 건강정보기술(health information system), 인터넷 활용 연구, EMR/EHR(Electronic Medical/Health Records)의 토픽 점유율은 증가하는 것으로 나타났다. 따라서 최근 의료 연구 분야에서는 데이터를 기반으로 의학적인 이해를 분석하고자 하는 연구들이 많이 수행되고 있다고 주장했다. Yang et al. (2018)은 스마트팩토리 분야의 연구 동향을 분석하기 위해 토픽 모델링 기법을 활용했다. 이 연구에서는 스마트팩토리 분야의 국내외 논문 텍스트 데이터를 수집하여 LSA (Latent Semantic Analysis) 기법을 통해 주제 및 관련 추세를 분석했다. 결론적으로 정보통신 기술, R&D, 기술 혁신 등의 주제에 대해서 국내 연구 및 해외 연구 동향을 비교 분석했다. 박한샘 외. 2019는 스마트 시티 관련 연구의 동향을 분석하기 위해 STM 기술을 활용했다. 이 연구는 Scopus에서 게재된 12,400건의 논문 데이터를 수집했다. 분류된 주제로는 '머신러닝', '네트워크 성능', '쓰레기 처리' 등 15개 토픽을 설정했다. 분석 결과 머신러닝 및 IoT(Internet of Things) 토픽이 증가하는 추세를 보였고, 소셜미디어 및 GIS(Geometry Information System) 토픽에서는 감소하는 추세를 보였다. Chen et al. (2020)은 Computer & Education 저널의 논문을 대상으로 STM 분석을 실시했다. 이 연구는 해당 저널의 논문 3,963건을 텍스트 데이터로 수집했다. 1976년부터 2018년까지의 연구 논문을 분석했으며, 주요 연구 토픽을 국가/지역별로 분석했다. 결론적으로 Computer & Education 저널의 투고를 위한 참고점을 제시했다.

Ⅲ. 제안 기법

3.1 DETM

LDA란 문서 집합을 분석하여 확률적으로



<그림 1> DETM Algorithm

주제를 생성하는 토픽 모델 기법을 의미한다(Blei et al., 2003). LDA 기법은 지금까지 기술 및 연구 동향 분석에서 문서의 숨겨진 주제를 분석하여 해당 분야의 대분류 및 트렌드를 분석하는 방식으로 활용되었다(Amado et al., 2018; Kim et al., 2018; Yang et al., 2018). 그러나 LDA 기법은 정해진 토픽 키워드만을 활용하기 때문에 특정 분야의 기술에 대한 트렌드를 분석할 수 없다는 한계점이 존재한다. 또한, LDA는 연구에서 활용되는 텍스트 데이터의 메타데이터(metadata)를 활용할 수 없다.

DETM(Dynamic Embedded Topic Model) 기법은 D-LDA(Dynamic Embedded Latent Dirichlet Allocation) 알고리즘을 기반으로 단어 임베딩(word embedding) 기술이 결합된 방식이다(Dieng et al., 2019). 여기서 D-LDA란 주제를 나타내는 다항 분포의 매개변수에 상태 공간(state-space) 모델 기법을 적용하여 각 주제별로 시계열적인 추세를 분석할 수 있는 방법론이다(Blei et al., 2006). 또한, 단어 임베딩은 단어 집합을 벡터 공간에 할당하는 방식으로 각 단어별 유사도를 측정할 수 있는 알고리즘을 의미한다. DETM 기법은 분석 초기 단계에서 랜덤 워크(random walk) 정의 시 RNN (Recurrent Neural Network) 기반 구조적 상각 변동 추론(amortized variational inference)을 적용하여 토픽 분류의 안정성을 보완했다.

<그림 1>은 DETM의 알고리즘 프로세스를 나타낸다(Dieng et al., 2019). 단어 토픽 분포 w_{dn}은 토픽 임베딩 $a_k^{(t)}$와 단어 임베딩 ρ_v의 내적 연산을 통해서 형성된다. $a_k^{(t)}$는 k번째 주제에 대한 단어 임베딩 벡터 공간에서의 분포 표현(distributed representation)으로 정의되며 ρ_v는 단어 임베딩의 벡터 표현값을 의미한다. w_{dn}은 단어 임베딩과 토픽 임베딩이 일치할 때 증가하기 때문에 유사한 단어는 유사한 주제에 할당된다. η_t는 시계열적으로 변화하는 잠재 변수(latent variables)를 의미한다. 시점 t에 따라서, 각 토픽에 속할 확률 θ_d는 η_t로 결정되며,

z_{dn}은 토픽을 결정하는 변수에 해당된다.

3.2 STM

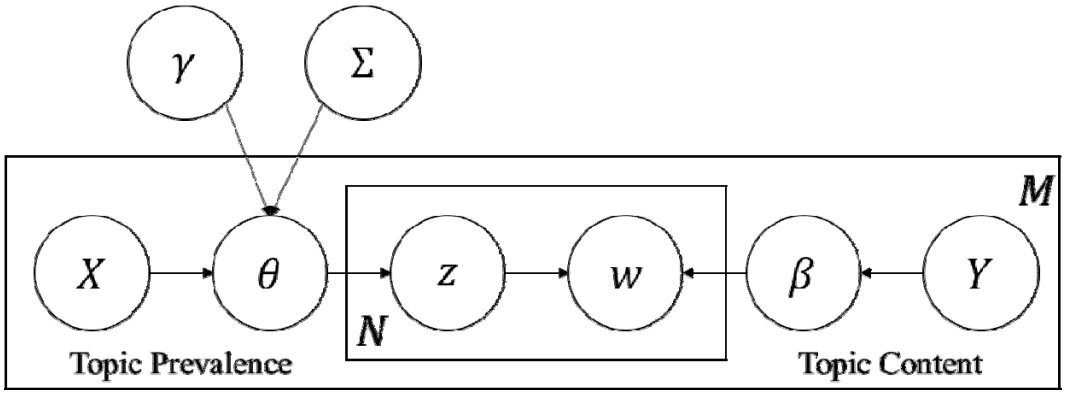
기존 LDA 기반 연구에서는 텍스트 데이터의 메타데이터를 제대로 활용할 수 없다는 한계점이 존재한다. 메타데이터를 활용할 수 있다면 각 토픽이 어떤 메타데이터의 속성에 가까운지 분석할 수 있다. 예를 들어 긍부정 레이블링이 존재하는 SNS 텍스트 데이터를 분석한다고 가정했을 때, 긍부정 정보를 메타데이터로 활용하면 각 토픽의 내용이 긍정적인지 부정적인지 나타낼 수 있다. STM(Structural Topic Model)은 이 점을 고려하여 텍스트 데이터의 메타 정보를 공변량으로 활용할 수 있도록 개발된 방법론이다(Roberts et al., 2013). STM 기법은 학습 과정에서 토픽 분류 결과를 개선할 뿐만 아니라, 메타데이터와 토픽간의 관계를 분석할 수 있다.

STM의 아키텍처는 <그림 2>와 같이 구성된다. 먼저 M은 말뭉치(corpus)의 문서 집합, N은 문서의 단어 집합을 의미한다. X는 Topic Prevalence를 생성하기 위해 사용되는 메타데이터를 의미하고 Y는 Topic Content를 생성하기 위한 메타데이터를 의미한다. Topic Prevalence는 θ로 표현되는데, 하나의 문서가 각 토픽에 속할 확률을 나타낸다. γ와 Σ는 각각 메타데이터의 가중치와 토픽의 공변량(topic covariances)를 의미하며, 토픽 확률 계산(θ) 시 반영된다. z는 θ로 결정된 토픽을 나타낸다. Topic Content는 문서에서 각 단어들의 토픽 확률값을 나타내는 것으로 β로 표현된다. w는 β를 기반으로 해당 토픽에서 선정된 단어를 의미한다.

Ⅳ. 분석

4.1 데이터

본 연구에서는 특허 텍스트 데이터를 수집하기 위해 특허정보 온라인 검색 서비스 WIPS ON을 활용했다. 해당 사이트에서 '섬유소재' 관련 키워드를 검색하여 1974년부터 2020년까



<그림 2> STM Algorithm

지 조회된 866건의 특허 데이터를 수집했다. 토픽 및 키워드에 대한 시계열 분석을 위해서 특허 문서 수가 2건 이하인 연도를 제외하고 1984년부터의 데이터를 활용했다. WIPS ON에서 제공하는 특허 데이터는 국가코드, 발명의 명칭, 요약문, 대표청구항, 인용 문헌 수 등 다양한 정보를 포함한다. 실제 분석에 적용된 특허 문헌의 상태 분류는 출원 13건, 심사 중 16건, 등록 423건, 거절 105건, 소멸 205건, 취하 32건, 포기 14건, 총 808건에 해당한다. 메타데이터 활용을 위한 레이블링(labeling)을 위해 상태 분류의 기준을 설정하여 각각 데이터를 수집 및 레이블링 작업 후 통합했다. 상태 분류 레이블링은 등록 여부에 따라서 분류했으며 출원 및 심사가 진행 중인 데이터는 제외하고 진행했다.

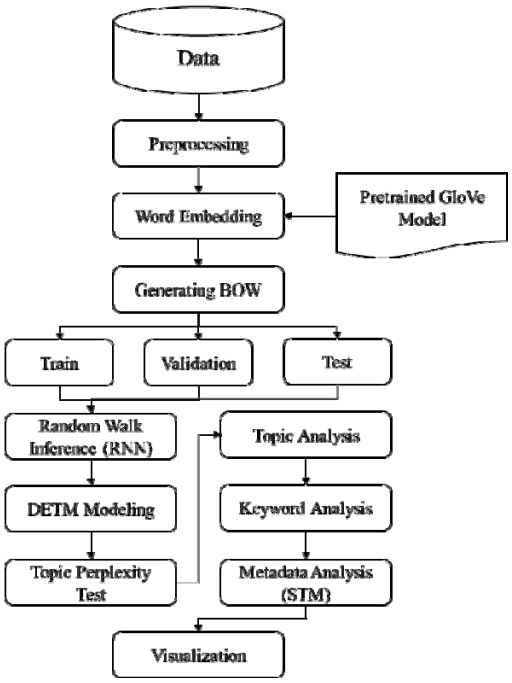
데이터 전처리의 경우 파이썬을 활용하여 각 특허 문서의 연도 정보를 추출 후 따로 칼럼을 구성했다. 텍스트 데이터는 특허 문서의 제목과 요약문을 통합하는 형태로 처리했으며, KoNLPy 패키지를 활용하여 POS(Part of Speech)를 태깅(tagging)하는 방식으로 토크나이징(tokenizing) 및 명사 추출 작업을 수행했다. 추출된 명사 토큰 중 의미가 없거나 품사가 잘못 분류된 경우 불용어 처리를 통해 제거했다. 데이터 정제 작업 후 각 명사 토큰에 대한 카운트 벡터(count vector)를 생성했다. 이 과정에서 2개 이하의 문서에서 출현하거나 70%의 문서에서 사용된 토큰들은 제거하는 과정을 거쳤다. 그 이유는 단어의 출현이 희소하거나 일상적인 용어의 경우 의미있는 토픽을 분류하는데 노이즈(noise)가 발생할 수 있기 때문이다. 다음 단계로 개별 단어에 대한 id를 부여하고 토픽 혼재성(topic perplexity)을 측정하기 위해 전체 데이터를 학습(train), 검증(validation), 시험(test) 데이터로 구분했다. 학습, 검증, 시험 데이터의 비율은 8.5:0.5:1로 할당했다. 토픽 혼재성이란 지표가 낮을수록 학습된 토픽 모델이 단어를 잘 예측한다는 의미로 해석된다. 최종적으로 문서별 단어 리스트를 추출하면서 BOW(Bag of Words)를 생성했다. 여기서 BOW는 단어들의 순서는 고려하지 않으면서 출현 빈도만 활용해서 텍스트 토큰을 수치화하는 방식을 의미한다.

4.2 모델링

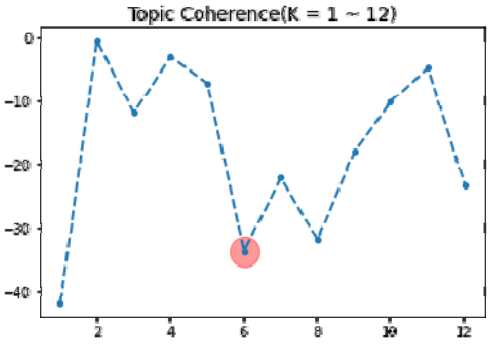
본 연구는 <그림 3>과 같이 DETM 방법론과 STM의 메타데이터 기반 메커니즘을 활용하여 토픽 모델링을 수행했다. DETM에서 단어 임베딩에는 GloVe(Global Vectors for Word Represenation) 기법을 적용했다. GloVe란 전역적 행렬 분해(global matrix factorization)와 지역적 context window의 장점을 결합한 전역적 대수 선형 회귀(global logbilinear regression) 모델을 의미한다(Pennington et al., 2014). 즉, 기존의 말뭉치의 전체적인 통계 정보를 고려할 수 있는 BOW와 예측 기반으로 임베딩 벡터를 생성하는 Word2Vec의 메커니즘을 통합한 것이다. 토픽 모델링 및 키워드 분석의 특성상 텍스트 데이터의 전반적인 통계와 문맥적 정보가 같이 고려되어야 하기 때문에 GloVe 기법을 선정했다. 본 연구의 GloVe 모델은 한국어 버전으로 공개된 사전학습 모델을 활용했다(ratsgo, 2019).

토픽 모델링 기법은 최적의 토픽 수 k를 토픽 일관성(topic coherence) 측정을 통해 결정한

다. 토픽 일관성이란 하나의 토픽에 포함되는 단어들의 상관관계를 의미하며 수치가 높을수록 토픽 내에서 유사한 단어가 많다는 것으로 해석된다. 토픽 일관성 수치가 과도하게 높으면 토픽의 키워드들이 단조로워지는 단점이 있고, 낮으면 제대로 토픽이 분류되지 않는 문제점이 발생한다. 본 연구에서는 <그림 4>와 같이 토픽 일관성을 고려하여 k를 6개로 설정했다. DETM 모델링 전 안정적인 토픽 분류를 위해 RNN 기반 상각 변동 추론을 활용하여 랜덤 워크를 정의했다. DETM 모델링 시 임베딩과 배치 사이즈는 100으로 설정했으며 학습률은 0.0001로, 학습 반복 횟수(epochs)는 1000번으로 설정했다. 모델링 학습의 경우 토픽의 혼재성이 가장 낮은 모델을 저장하는 방식으로 최적화했다. 토픽 모델링 결과를 바탕으로 각 토픽별로 베타(β)값이 높은 키워드들을 추출했다. DETM 결과의 토픽 점유율을 구하기 위해서 각 문서에 대한 토픽별 세타(θ)값을 계산하여 평균을 구하는 방식으로 토픽 점유율을 계산했다. 본 연구에서는 토픽 점유율을 활용하여 연도별 토픽 트렌드를 분석했다. 키워드 분석의 경우 해당 키워드와 유사한 단어 10개를 워드 임베딩을 통해서 추출 후 베타값을 더하는 방식으로 트렌드를 분석했다. 메타데이터 분석에서는 STM 기법의 기본 메커니즘을 활용하여 특허 문서의 등록 여부를 메타데이터로 설정했다. 메타데이터를 기반으로 각 토픽의 메타 효과(meta effect)를 측정했으며, 메타데이터별 토픽 트렌드를 분석했다. 메타 효과는 각 문서의 레이블링 정보를 활용하여 등록된 특허 문서의 세타값 평균에서 비등록 특허 문서의 세타값 평균을 빼는 방식으로 계산되었다.



<그림 3> Analysis Process



<그림 4> 토픽 개수별 Topic Coherence

Ⅴ. 연구 결과

5.1 토픽 분석 결과

본 연구의 토픽 분석 결과는 다음과 같다. 6

<표 1> 토픽 분석 결과

id	토픽명	토픽 점유율	키워드
Topic 1	공정	0.08408	열기, 체인, 마찰, 나노, 포장, 스마트, 기구, 작동, 기능 등
Topic 2	제조	0.52563	제조, 원단, 피혁, 중량, 분리, 공정, 복합, 직물, 코팅, 염료 등
Topic 3	기능성 의복	0.03863	방적, 속옷, 물질, 세라믹, 발가락, 림프, 보정속옷, 발열, 신발 등
Topic 4	상품	0.13678	커튼, 레이스, 방지, 고무, 염료, 신발, 함유, 장치, 수축 등
Topic 5	탄성	0.21426	모양, 인조섬유, 탄성, 폴리에스테르, 라텍스, 충격, 자동차, 질감 등
Topic 6	안전	0.00063	쿠션, 제어장치, 나노, 구성, 고무, 탄성, 부착, 기능, 받침, 시트, 방지 등

개의 토픽에 대해서 <표 1>과 같이 해당 키워드들을 활용하여 토픽명을 설정했다. Topic 1은 공정이라는 주제로 열기, 체인, 마찰 등의 키워드가 포함된다. 섬유소재의 공정 분야에서는 스마트 및 나노 기술에 대한 관심이 나타났다. Topic 2는 제조 관련 주제로 제조, 원단, 피혁 등의 키워드가 나타났다. 제조 관련 토픽은 점유율이 0.52563으로 가장 높게 나타났다. 토픽 점유율 수치의 의미는 Topic2의 경우 전체 문서에서 약 52%가 제조 토픽으로 분류되었다는 것이다. Topic 3은 기능성 의복의 주제로 방적, 속옷, 물질 등의 키워드가 포함된다. Topic 3을 통해 섬유소재 분야에서 보정속옷, 신발 등 다양한 물품에 대한 특허 기술이 존재한다는 것을 파악했다. Topic 4는 상품관련 토픽으로 커튼, 레이스, 방지 등의 키워드가 나타났다. Topic 5의 경우 모양, 인조 섬유, 탄성 등의 키워드로 탄성에 관련된 주제로 분석되었다. Topic 5는 토픽 점유율이 0.21426으로 섬유소재 분야에서 탄성에 관련된 특허 문서들이 많다는 것이 확인되었다. Topic 6은 점유율이 가장 낮은 토픽으로 안전을 위한 섬유소재를 연구하는 특허 주제로 분석되었다.

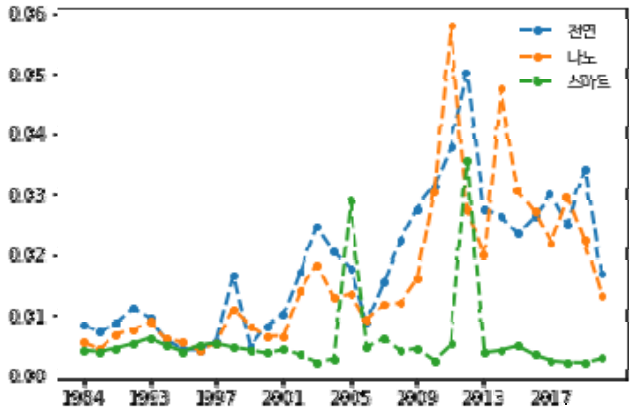
토픽 트렌드 분석 결과는 <그림 5>와 같이 나타났다. 토픽 트렌드 분석 결과는 시간적인 흐름에 따라서 각 토픽과 관련된 문서들 수의 증감을 나타낸다. <그림 5>의 파란선은 각 토픽의 점유율, 주황색 점선은 전반적인 추세선을 의미한다. Topic 1의 경우 토픽 점유율이 연도별로 증가하는 추세를 보인다. Topic 1의 트렌드 분석 결과를 통해 섬유소재 분야에서 공정에 관련된 주제로 많은 특허 연구들이 진행된 것으로 분석된다. 또한, <그림 5>의 그래프를 분석했을 때, 가장 최근 연도인 2019년에 크게 증가된 것으로 측정되기 때문에 공정 관련 나노 및 스마트 기술에 대한 관심이 늘어나고 있는 것이 확인되었다. Topic 2의 트렌드 분석 결과 제조 관련 기술은 꾸준하게 점유율을 가지고 있는 것으로 분석되었다. Topic 3의 결과는 기능성 의복 관련 기술이 2000년대 중반부터 큰 관심을 받다가 최근에는 점유율이 크게 감소된 양상이 나타났다. Topic 4의 토픽 트렌드는 점차적인 상승 곡선이 나타났으며 특히 2019년 토픽 점유율이 크게 나타난 것으로 봤을 때, 앞으로도 섬유소재 분야의 상품 관련 특허 기술들이 활발하게 연구될 것으로 분석되었다. Topic 5의 경우 큰 변동없이 트렌드가 나타난 것으로 봤을 때 섬유소재의 탄성에 대한 연구는 꾸준하게 수행되고 있는 것으로 확인되었다. Topic 6의 트렌드는 2004년과 2011년 큰

<그림 5> 토픽 트렌드 분석 결과

토픽 점유율이 측정되지만 최근에는 낮은 수치로 분석되었다.

5.2 키워드 분석 결과

본 연구의 키워드 분석 결과는 <그림 6>과 같다. 연구자가 임의로 설정한 키워드의 토픽 점유율을 연도별로 나타낸 것이다. 토픽 키워드의 경우 최근 섬유소재 분야에서 주목받고 있는 '천연', '나노', '스마트'를 입력값으로 설정했다(대전일보, 2021; 한국섬유산업연합회, 2021). 각 키워드의 점유율 수치는 단어 임베딩을 통해서 추출된 유사 단어 10개의 해당 값이 더해진 것이다. 즉, 본 연구의 키워드 분석 결과는 하나의 키워드를 분석하는 것보다 더욱 넓은 의미적 영역을 포함한다. 키워드 설정은 섬유소재 분야에서 일반적으로 사용되는 용어로 선정했다. 최근 연도에서 키워드들이 전반적으로 하락하는 추세를 보이는데, 이는 섬유 소재 분야에서 최근 다양한 주제들이 등장하여 상대적으로 키워드의 점유율이 낮아지는 것으로 분



<그림 6> 키워드 트렌드 분석 결과

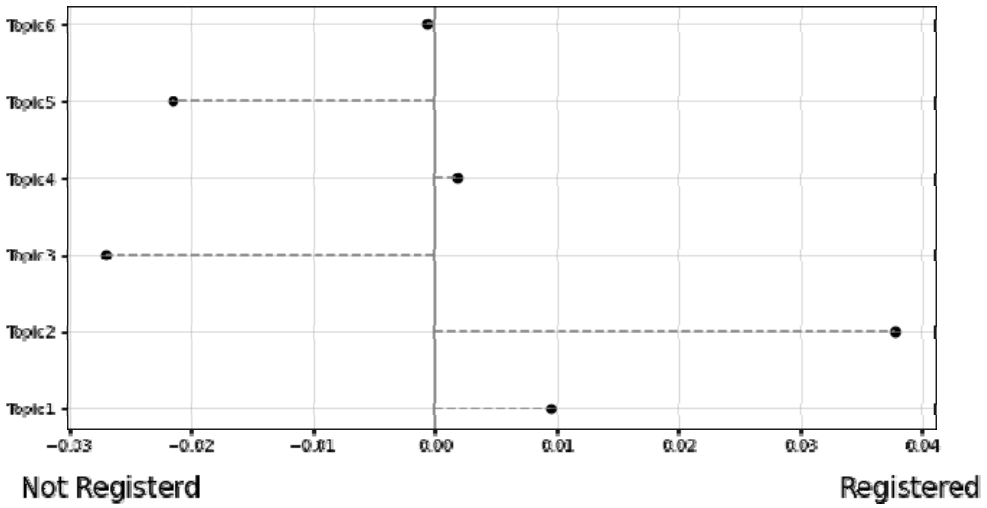
석되었다.

'천연' 키워드의 경우 시계열 전반적으로 높은 수치와 상승하는 추세가 나타났다. 이는 섬유 소재 분야의 연구에서 천연 기술에 대한 관심이 꾸준하게 증가했음을 의미한다. 천연 섬유 기술은 최근 재활용 및 향균 기술과 함께 주목받고 있는 기술이다(한국섬유산업연합회, 2021). '나노' 키워드는 천연 섬유 기술과 같이 점차적으로 높은 수치로 측정되었다. 특히 2000년대 후반에 키워드 중 가장 큰 상승 추세를 보였다. 나노 섬유 분야는 최근 친환경 소재인 나노 셀룰로오스 기술이 주목받고 있는 상황이다(대전일보, 2021). 나노 셀룰로오스 소재는 식물 섬유에서 추출하는 나노미터 크기의 천연 고분자 물질로써 가벼우면서 가공이 쉽고, 투명성, 생체 안정성, 생분해성 등 다양한 장점을 가지고 있다. '스마트' 키워드는 2005년과 2012년에 높은 점유율이 측정되었지만 다른 연도에서는 상당히 낮은 수치로 나타났다. 이는 최근 스마트 섬유 기술이 국내외 섬유소재 분야에서 주목받고 있지만, 아직 국내 특허에서는 관심도가 부족하다는 점을 시사한다.

5.3 메타데이터 분석 결과

STM의 기본 메커니즘을 활용한 메타데이터 분석 결과는 <그림 7>과 같다. 그래프의 메타 효과 수치는 각 토픽이 특허 등록 여부에 얼마나 영향을 주는가를 나타낸다. 즉, 메타 효과의 수치가 클수록 해당 토픽이 특허 등록과 관계가 있다는 것이다. 따라서 <그림 7>의 메타 효과의 수치는 특허 문서의 등록 가능성과도 관계가 있다고 볼 수 있다. 등록 여부 레이블링 정보를 활용할 수 있는 특허가 한정적이기 때문에, 출원 및 심사 중인 특허를 제외하고 2004년부터 2019년까지의 특허 문서에 대한 메타 효과를 측정했다.

Topic 1과 2에서는 메타 효과의 수치가 양수로 나타났다. 이는 섬유소재 분야에서 공정과 제조 관련 특허 문서들이 등록이 된 경우가 많은 것으로 분석되었다. 특히 Topic 2은 가장 높은 메타 효과 수치가 나타났는데, 이는 특허 문서 수가 가장 많으면서 등록 가능성이 높은 것으로 섬유소재 분야에서 제조 기술이 주요 분야라는 점을 시사한다. Topic 3, 5는 메타 효과



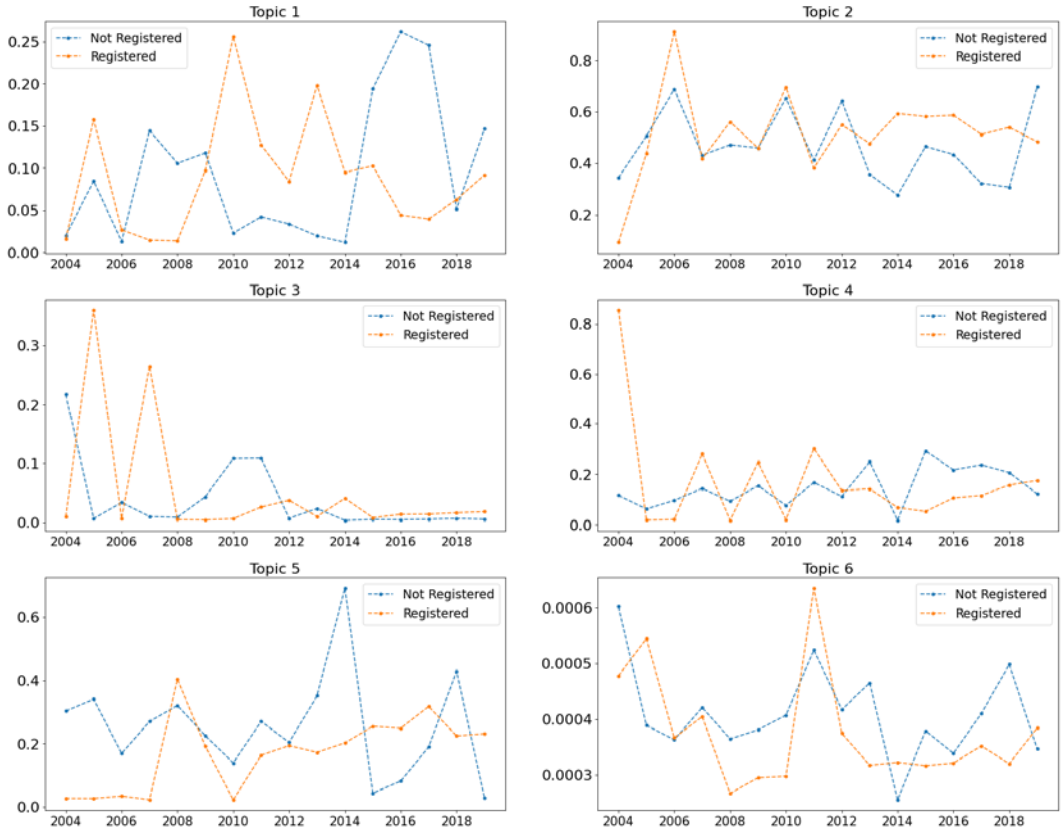
<그림 7> 토픽별 메타 효과

의 수치가 음수로 측정되었다. 이는 섬유소재 분야에서 전반적으로 기능성 의복과 탄성에 관련된 많은 연구 시도가 있었지만, 특허 등록에 어려움이 있었던 것으로 분석되었다. Topic 4, 6에 해당하는 상품과 안전 주제의 특허들은 메타 효과가 크게 드러나지 않는 것으로 나타났다.

특허 문서의 전반적인 메타 효과는 <그림 7>의 결과에서 확인할 수 있지만, 시계열적인 추세 분석에서는 다른 결과가 도출될 수도 있다. 특히 특허 기술의 경우 최근 트렌드가 중요하기 때문에 최근 연도의 메타 효과를 분석하는 것이 중요하다. <그림 8>의 분석 결과는 시계열적인 메타 효과를 측정한 것이다. Topic 1의 경우 2010년대 초반에서 등록된 특허 문서들의 세타값이 큰 수치로 측정되었지만 최근 연도에서는 크게 떨어지는 것으로 나타났다. 이는 공정 관련 특허 문서들이 최근에 등록이 되지 않는 사례들이 발생했다는 것을 의미한다. Topic 2의 경우 2013년 이후 등록된 특허 문서들의 세타값이 증가했으나 2019년에는 비등록 특허 문서들의 점유율이 증가한 것으로 나타났다. Topic 1, 2는 전반적인 메타 효과는 양수로 측정되었지만 최근 연도에서 수치가 하락하는 추세로 나타났다. 반면에 Topic 5, 6은 2010년대 초반에 메타 효과 값이 낮아지는 것으로 보이지만 2019년에서 등록 특허 문서의 세타값이 오르는 추세가 확인되었다. 결과적으로 최근 연도의 토픽별 메타 효과를 분석했을 때 전반적인 효과와 상반되는 결과가 나타났다. 특히 <그림 7>과 같이 전체 토픽의 메타 효과를 분석할 때 수치가 낮은 경우 시계열 추세 분석이 필수적이라는 점을 시사한다. Topic 3, 4는 시계열적인 추세를 분석했을 때 최근에 등록 및 비등록 특허 문서의 세타값이 큰 차이가 없는 것으로 측정되었다.

Ⅵ. 결 론

본 연구는 DETM & STM 텍스트마이닝 방



<그림 8> 토픽별 메타 효과 추세

법론을 활용하여 섬유소재 분야의 특허 기술 동향을 분석했다. 분석 결과 분류된 키워드를 기반으로 '공정, 제조, 기능성 의복, 상품, 탄성, 안전'의 6개 토픽명을 설정했다. 전체 토픽 점유율 분석에서는 Topic 2에 해당하는 제조 관련 특허 문서들이 가장 많은 것으로 나타났다. 토픽 트렌드 분석에서는 공정 및 상품 관련 특허 문서의 점유율이 증가하는 추세를 보였다. 키워드 트렌드 분석에서는 천연 및 나노 섬유 기술이 최근 특허 기술로서 주목받고 있다는 사실을 확인했으며, 스마트 관련 연구는 아직 국내에서 부족하다는 결론을 내렸다. STM 메커니즘 기반의 메타데이터 분석 결과 공정과 제조 분야에서 메타 효과가 큰 수치를 나타냈다. 즉, 섬유소재 분야에서 공정과 제조 관련 특허 기술이 등록 가능성이 높을 수 있다는 결론을 도출했다. 하지만 메타 효과의 시계열 추세에서는 상반된 결과로 공정과 제조 관련 특허 문서에서 메타 효과의 수치가 낮아지고 탄성과 안전 기술에 대한 메타 효과의 수치가 2019년에 증가하는 추세로 나타났다. 즉, 섬유소재 분야에서의 전반적인 특허 기술 동향은 최근 섬유 공정 및 상품 관련 특허 출원이 증가하고 있으며, 천연 및 나노 기술이 주목받고 있고, 탄성과 안전 관련 특허 문서가 등록 가능성이 높다고 분석되었다.

결론적으로 본 연구의 시사점은 다음과 같다. 첫 번째, 국내 섬유소재 분야의 연구자 및 실무자들에게 기술적 참고점을 제시했다. DETM & STM 기법을 활용하여 실증적으로 국내 특허 기술의 동향을 다각적으로 분석했다. 두 번째, 연구 및 기술 동향 분석의 새로운 방향성을 제시할 수 있는 프레임워크를 개발했다. <그림 3>의 모델링 프레임워크는 DETM & STM 방법론을 기반으로 새롭게 개발된 방식이다. 본 연구의 방법론은 시계열 및 메타 정보가 있는 텍스트 데이터를 대상으로 심층적인 분석이 가능하다. 또한, 단순 LDA 기반의 토픽 모델링 분석에서는 결과를 산출하기 어려운 메타데이터별 트렌드 및 키워드 트렌드 분석이 가능하다. 특히 DETM 방법론의 경우 토픽 모델링 과정에서 단어 임베딩 기법이 함께 적용되기 때문에 더욱 정밀한 키워드 트렌드 분석을 할 수 있다. 세 번째, 본 연구는 거시적 관점에서 전체 토픽 및 메타 효과를 분석하고, 미시적 관점에서는 각 토픽 및 키워드에 대한 시계열 분석을 통해 전반적인 특허 기술 동향을 제시했다. 메타 효과 분석 결과에서는 전반적으로 공정 및 제조 기술의 특허 문서에서 등록된 비율이 높았지만, 최근 연도의 추세에서는 오히려 탄성과 안전 관련 기술들이 등록 메타 효과가 증가하는 추세를 확인했다. 본 연구의 결과 및 방법론을 활용하여 섬유소재 분야의 연구와 해당 산업 분야의 활성화를 기대한다. 네 번째, 본 연구의 분석 방법론을 활용하여 섬유소재 산업의 기술자들을 위한 자동 분석 보고서 서비스를 개발할 수 있다. 기존 연구에서도 특정 기술에 대한 자동 분석에 대한 수요 및 필요성을 제시했다(Tseng et al., 2007). 본 연구의 방법론은 토픽 트렌드, 키워드, 메타 효과 분석이 가능하기 때문에 해당 분야의 전반적인 분석이 가능하다. 특허 문서뿐만 아니라 논문, 신문 기사 등의 데이터를 같이 활용한다면 높은 수준의 정보를 추출할 수 있는 자동 분석 보고서 서비스로 활용할 수 있다.

본 연구의 한계점은 적은 수의 국내 특허 데이터를 활용했다는 점이다. 연구 및 기술 동향 분석에서는 텍스트 데이터가 많을수록 유용하고 의미 있는 결과를 도출할 수 있다. 또한, STM 메커니즘을 활용하여 메타 효과를 분석했으나 학습 과정에는 메타데이터가 활용되지 않았다. 본 연구의 제안 기법에서 STM의 메커니즘을 기반으로 메타데이터를 활용한다면 토픽 모델링 결과의 개선이 기대된다. 향후 연구에서는 국내 데이터뿐만 아니라 국외 데이터를 추가하고, 메타데이터를 학습 과정에서 활용한다면 더욱 고도화된 연구를 수행할 수 있을 것이다.

참고문헌

김민수, 김희웅, “온라인 해킹 불법 시장 분석: 데이터 마이닝과 소셜 네트워크 분석 활용”, 정보시스템연구, 제29권, 제2호, 2020, pp. 221-242.

매일신문, “다이텍연구원, ‘비대면 섬유소재 마케팅 플랫폼’ 만든다”, 신중언 기고, 2020.12.29.

대전일보, “[숲 사랑] 친환경 미래 소재, 나노 셀룰로스”, 박현 기고, 2021.08.31.

박한샘, 김동현, 장성주, “Research Trend Analysis on Smart City Based on

Structural Topic Modeling (STM)", 디지털콘텐츠학회논문지 (J. DCS), 제20권, 제9호, 2019, pp. 1839-1846.

최병설, 김남규, "감정 딥러닝 필터를 활용한 토픽 모델링 방법론", 정보시스템연구, 제28권, 제4호, 2019, pp. 271-291.

한국섬유산업연합회, "프리뷰 인 서울 2021 프레스 킷", 2021.

한국전자통신연구원(ETRI), "빅데이터 플랫폼의 산업생태계 현황과 주요 이슈", 김문구, 박종현, 2019, pp. 7-17.

홍태호, 임강, 박지영, "Lda를 이용한 온라인 리뷰의 다중 토픽별 감성분석-Tripadvisor 사례를 중심으로", 정보시스템연구 제27권, 제1호, 2018, pp. 89-110.

Amado, Alexandra., Paulo Cortez., Paulo Rita., Sérgio Moro., "Research Trends on Big Data in Marketing: A Text Mining and Topic Modeling Based Literature Analysis," *European Research on Management and Business Economics*, Vol. 24, No. 1, 2018, pp. 1-7.

Blei, David M., John D Lafferty., "Dynamic Topic Models," *Paper presented at the Proceedings of the 23rd International Conference on Machine Learning*, 2006, pp 113-120.

Blei, David M., Andrew Y Ng., Michael I Jordan., "Latent Dirichlet Allocation," t*he Journal of Machine Learning Research*, No. 3, 2003, pp. 993-1022.

Chen, X., Zou, D., Cheng, G., & Xie, H., "Detecting Latent Topics and Trends in Educational Technologies over Four Decades Using Structural Topic Modeling: A Retrospective of All Volumes of Computers & Education." *Computers & Education*, No. 151, 2020.

Dieng, Adji B., Francisco JR Ruiz., David M Blei., "The Dynamic Embedded Topic Model," *arXiv preprint arXiv: 1907. 05545*, 2019.

Hofmann, Thomas., "Probabilistic Latent Semantic Indexing," *Paper Presented at the Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999, pp. 50-57.

Kho, Jaechang., Kuentae Cho., and Yoonho Cho., "A Study on Recent Research Trend in Management of Technology Using Keywords Network Analysis," *Journal of Intelligence and Information Systems*, Vol. 19, No. 2, 2013, pp. 101-123.

Kim, Yong-Mi., Dursun Delen., "Medical Informatics Research Trend Analysis: A Text Mining Approach," *Health Informatics Journal*, Vol. 24, No. 4, 2018, pp. 432-452.

Noh, Heeyong., Yeongran Jo., and Sungjoo Lee., "Keyword Selection and Processing Strategy for Applying Text Mining to Patent Analysis," *Expert Systems with Applications*, Vol. 42, No. 9, 2015, pp. 4348-4360.

Pennington, Jeffrey., Richard Socher., Christopher D Manning., "Glove: Global

Vectors for Word Representation," *Paper Presented at the Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532-1543.

Ratsgo, "Korean Word Embedding Pre-trained Model", https://github.com/ratsgo/embedding/releases , 2019.

Roberts, Margaret E., Brandon M Stewart., Dustin Tingley., and Edoardo M Airoldi., "The Structural Topic Model and Applied Social Science," *Paper Presented at the Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*, Vol 4, 2013, pp 1-20.

Scott, John., "Social Network Analysis," *Sociology*, Vol. 22, No. 1, 1988, pp. 109-127.

Tseng, Yuen-Hsien., Chi-Jen Lin., and Yu-I, Lin., "Text Mining Techniques for Patent Analysis," I*nformation Processing & Management*, Vol. 43, No. 5, 2007, pp. 1216-1247.

Yang, Hyun-Lim., Tai-Woo Chang., Yerim Choi., "Exploring the Research Trend of Smart Factory with Topic Modeling," *Sustainability*, Vol. 10, No. 8, 2018, pp. 2779.

Yoon, B., Park, Y., "A Text-mining-based Patent Network: Analytical Tool for High-technology Trend," *The Journal of High Technology Management Research*, Vol. 15, No. 1, 2004, pp. 37-50.

이 현 상 (Lee, Hyun Sang)

경북대학교 경영학사와 석사학위를 취득하였다. 현재 경북대학교 박사과정을 하고 있으며, 주요 관심분야는 경영정보, 딥러닝, 머신 러닝, 자연어 처리, 시계열 예측 등이다.

조 보 근 (Jo, Bo Geun)

미국 Eastern Michigan University에서 Computer Information System 전공 학사학위를 취득하고, 경북대학교 경영학부에서 석사학위를 취득하였다. 현재 경북대학교에서 경영학부 박사과정 중에 있으며, (주)빅웨이브에이아이에 이사로 재직 중이다. 주요 연구 분야는 기계학습, 딥러닝, 텍스트 마이닝 등이다.

오 세 환(Oh, Se Hwan)

현재 경북대학교 경영학부에서 부교수로 재직 중이다. 서울대학교 경제학부(학사)를 졸업했으며 카네기멜론대에서 e-비즈니스 석사, 서울대학교에서 경영학(경영정보) 박사학위를 받았다. 주요 연구 관심분야는 전자상거래, 국제무역에서의 IT 활용 등이다.

하 성 호 (Ha, Sung Ho)

연세대학교 경영학사와 한국과학기술원 석사와 박사학위를 취득하였다. 현재 경북대학교 경영학부 교수로 재직하고 있으며, 주요 관심 분야는 경영 분야에 적용되는 데이터 사이언스, 기계학습, 딥러닝 등이다.

<Abstract>

Research of Patent Technology Trends in Textile Materials: Text Mining Methodology Using DETM & STM

Lee, Hyun Sang · Jo, Bo Geun · Oh, Se Hwan · Ha, Sung Ho

Purpose

The purpose of this study is to analyze the trend of patent technology in textile materials using text mining methodology based on Dynamic Embedded Topic Model and Structural Topic Model. It is expected that this study will have positive impact on revitalizing and developing textile materials industry as finding out technology trends.

Design/methodology/approach

The data used in this study is 866 domestic patent text data in textile material from 1974 to 2020. In order to analyze technology trends from various aspect, Dynamic Embedded Topic Model and Structural Topic Model mechanism were used. The word embedding technique used in DETM is the GloVe technique. For Stable learning of topic modeling, amortized variational inference was performed based on the Recurrent Neural Network.

Findings

As a result of this analysis, it was found that ‘manufacture’ topics had the largest share among the six topics. Keyword trend analysis found the fact that natural and nanotechnology have recently been attracting attention. The metadata analysis results showed that manufacture technologies could have a high probability of patent registration in entire time series, but the analysis results in recent years showed that the trend of elasticity and safety technology is increasing.

Keyword: Textile Materials, Patent Text Data, Technique Trend Analysis, Dynamic Embedded Topic Model, Structural Topic Model

* 이 논문은 2021년 8월 31일 접수, 2021년 9월 17일 1차 심사, 2021년 9월 29일 게재 확정되었습니다.