기술경영학석사학위논문

기술 테마 분석을 통한 국내 조선산업의 친환경 기술 트렌드 분석

2022년 2월

부 경 대 학 교 대 학 원

기술경영학과

민 건 태

기술경영학석사학위논문

기술 테마 분석을 통한 국내 조선산업의 친환경 기술 트렌드 분석

지도교수 서 원 철

이 논문을 기술경영학석사 학위논문으로 제출함.

2022년 2월

부 경 대 학 교 대 학 원

기술경영학과

민 건 태

민건태의 기술경영학석사 학위논문을
인준함

2022년 2월 25일

위 원 장 경제학박사 이 민 규 (인)

위 원 공 학 박 사 최 성 철 (인)

위 원 공 학 박 사 서 원 철 (인)

목 차

표 목 차

그 림 목 차

기술 테마 분석을 통한 국내 조선산업의 친환경 기술 트렌드 분석

민 건 태

부 경 대 학 교 기 술 경 영 전 문 대 학 원

요 약

잠재 디리클레 할당(LDA)은 대량의 문서에 포함된 단어의 확률적 관계를 파악해 키워드를 표현하는 자연어 처리 알고리즘 가운데 하나다. 하나의 특성에 모인 키워드를 보고 토픽을 추론하는 형태로 분석이 진행된다. LDA의 이런 특성으로 말미암아 특허 분석에도 LDA를 활용한 연구가 활발히 진행되고 있다. 특허의 서지 정보를 분석뿐 아니라 요약문에 포함된 기술의 특성을 언어로 표현이 가능하므로, 특정 산업 내에서의 기술의 내적 특성을 파악할 수 있다는 장점이 있다.

이 연구는 국내 조선산업의 특허 기술을 LDA를 활용해 분석했다. 조선산업에 LDA를 적용한 이유는 선박 건조 공정을 중심으로 한 혁신의 과정이 최근 급격하게 변화했기 때문이다. 선박의 배기가스를 줄이기 위한 장치나 선박 평형수 내 미생물을 처리하는 설비가 대표적인 사례다. 엔진 역시 디젤과 LNG 활용 등 이중 연료 사용부터 시작해 최근에는 수소 추진 선박을 개발하기 위한 움직임이 나타난다. 선박의 부품과 장치에서 시작돼 선박 자체가 친환경 영역으로 바뀌는 추세다.

따라서 지난 10년 동안의 조선산업 부문에 대한 대량의 특허 문서를 수집해 기술적 특성을 파악했다. 한국조선해양기자재공업협동조합 회원사 255개사의 10년치 특허 13,607건을 분석했다. 이 특허 문서에서 도출한 64개의 토픽은 중소벤처기업부가 발표하는 '중소기업 전략기술로드맵 2020-2023'에 따라 8개의 카테고리 기준에 맞춰 분류했다. 전략기술로드맵에서 제시한 8개 카테고리(배기가스 저감 장치, LNG 선박용 기자재, 듀얼 추진 선박용 기자재, 복합경량소재 적용 중소형 선박,

레저선박 기자재 및 레저기기, 오염물 제거 시스템, 선박충돌 감지 시스템, 기관실용 기자재)가 각각의 기술 테마를 나타낸다.

연구 결과 8개 기술 테마에 64개의 토픽 중 47개가 해당되는 것으로 나타났다. 특허와 토픽을 연결한 매트릭스를 만든 뒤, 이를 다시 '친환경 기술'과 나머지 기술로 분류해 연도별 특허량을 비교했다. 양쪽 모두 2012년을 정점으로 감소하는 추세를 보였지만, 양적 측면에서는 친환경 관련 기술 개발이 압도적으로 많이 이뤄졌음을 밝혀냈다.

특허의 서지 정보를 통한 R&D 활동의 양적 분석에서 벗어나 기술의 특성을 포괄적으로 살펴봤다. 분석 과정에서 해양플랜트 관련 기술에 관한 토픽도 다수 발견됐는데, 마찬가지로 2012년을 기점으로 관련 기술이 대폭 줄었다는 사실을 확인했다. IMO 선박 규제와 관련한 친환경 선박 관련 기술개발 활동이 활발하게 이어졌음을 알 수 있다. 나아가 해양플랜트와 친환경 관련 기술의 연관성, 수소추진선박과 자율운항선박의 개발 가능성을 위한 기술 특성을 파악하는 연구로 이어질 것으로 기대된다.

Ⅰ. 서 론

2000년대 중반 이후 중국의 위협을 받던 한국의 조선산업이 최근 재평가 받고 있다. 국내 조선사는 2020년 수주량을 기준으로 세계 1위를 달성했으며, 2021년(1~8월 기준) 역시 13년 만에 가장 많은 수준의 수주량을 기록하며 주도권 경쟁에서 우위를 확보했다.[1)2)]

긴 침체기를 겪은 한국의 조선산업이 최근 두각을 보인 것은 국제해사기구(IMO)의 선박 환경 규제와 맞물려 친환경 선박이라는 새로운 수요가 창출됐기 때문이다.

양해성 등 (2021)은 조선산업 부문에서의 주도권 확보 과정을 산업의 추격 차원에서 분석했다. 조선산업이 중저기술 영역에 속해 있는 상황에서 기술적 기회의 창은 제한적일 수밖에 없지만, 국제적인 선박 오염 규제가 적용되면서 새로운 시장이 창출되고, 이에 따라 규제·수요적 기회의 창이 생겼다는 설명이다. 선박이 배출하는 황산화물(SOx)과 질소산화물(NOx)에 관한 규제가 석유에서 천연가스로 동력원의 전환을 유도하고, 특정 해역에 존재하는 미생물의 이동을 막기 위해 선박평형수 처리장치라는 새로운 기술을 탄생시킨 계기가 된 셈이다.

선박 제조 공정에서 원가 절감을 목표로 이뤄진 모듈화 등의 점진적 기술혁신이 그동안의 기술 트렌드였다면, 앞으로는 국제적 규모의 환경 규제를 중심으로 한 친환경 기술이 조선산업의 새로운 기술 트렌드로 자리매김할 것으로 예상된다. IMO는 2000년대 중반부터 규제안을 내놓은 뒤 단계적으로 규제를 강화해 최종적으로

---

1) 산업통상자원부가 2021년 1월 발표한 조선업 수주 물량에서 한국은 세계 1위를 기록했다. 2020년 하반기 LNG 운반선 등 고부가가치 선박 부문에서 6,840,000 CGT 규모의 수주를 달성해 중국을 따돌린 것으로 조사됐다.

2) BNK 경제연구소가 발표한 조사에서도 2008년 이후 국내 조선산업이 13년 만에 최고 실적을 기록했다는 조사 결과를 발표했다. 컨테이너선과 LNG선을 중심으로 수주 실적이 개선됐으며, IMO 환경 규제 영향 속에서 국내 기업이 적극적으로 기술을 개발한 결과라고 평가했다.

이산화탄소 배출이 없는 선박이 바다를 누리는 방안을 구상 중이다.

이 연구의 목적은 국내 조선사와 선박용 기자재를 개발하는 중소기업을 대상으로 특허 기반의 기술 트렌드를 살펴보며 국내 조선산업의 기술 진화의 형태를 분석하는 데 있다.

서원철(2021)은 특허가 기술혁신 분석을 위한 중요한 도구이므로, 특허정보 기반의 기술기회 발굴 관련 연구가 다양한 관점에서 진행된다고 설명했다.

Yoon 등 (2004)은 특허 문서 자체가 기술적 진화 과정의 중요한 기초 지식을 포함하기 때문에 특허 정보를 중심으로 다양한 형태의 분석이 진행된다고 봤다.

Zhang 등 (2017)은 전기차 배터리 관련 기술을 지역 등 카테고리 별로 분류해 이에 대한 발명의 양적 성장률을 측정하고 나아가 발명인 간의 네트워크를 분석한 뒤 미래의 주요 기술 분야를 예측했다.

조용래 등 (2014)은 특허 서지정보를 국제산업분류체계에 따라 분류한 다음 네트워크 분석 기법을 적용해 기술융합의 매커니즘을 규명하는 연구를 진행했다.

서지 정보를 활용해 새로운 기술을 발견하거나, 제품 기회를 분석하는 연구도 이뤄졌다. Shin과 Seo(2017)는 IPC에 기반한 정보를 토대로 기업의 내적 역량을 파악해 새로운 기술 영역을 정의한 뒤 해당 산업과의 포괄적 연관성을 분석하는 방법론을 적용했다.

하지만 기술분류코드나 인용 정보 등의 서지정보만을 활용한 분석은 특허에 내포된 기술적 의미를 파악하는 데 한계가 있다. 이에 따라 특허 문서에서 단어 행렬을 추출, 다양한 분석 기법을 적용해 기술기회 또는 공백기술을 예측하는 연구가 이뤄지고 있다.

전성해(2011)는 특허 서지 정보에서 벗어나, 특허 문서를 기반으로 앙상블 모델을 적용해 공백 기술을 예측하는 연구를 진행했다.

이지호 등 (2018)은 자동차 부품 기업을 중심으로 특허 문서를 수집해 연구했다. 특허로부터 기술적 문제와 해결방법을 각각 묶어 이를 네트워크로 표현한 뒤, 기술

경쟁의 동향을 파악하는 방법론이다.

Korobkin 등 (2018)은 특허 문서를 수집한 뒤 SAO(Subject-Action-Object) 구조를 도출해 특허 간의 의미론적 유사성을 파악하는 시도를 했다.

강지호 등 (2015)은 CPC를 활용해 연관규칙 마이닝 기법을 적용, 웨어러블 기술 분야에서의 기술융합 동향을 파악했다.

잠재 디리클레 할당(LDA) 모델은 특허 문서를 토대로 한 대표적인 분석 기법이다. Jeong과 Yoon(2017)은 LDA 모델링이 대량의 특허 문서를 분석해 단어와 문서 간의 확률적 연관성에 따라 키워드를 나열해 토픽을 추론하는 형태로, 특허와 산업에 관해 포괄적으로 이해할 수 있어 개별 기업의 R&D 기획의 방향은 물론 기술의 진화 방향을 파악이 가능하다고 설명했다.

본 논문에서는 한국의 조선사와 선박용 기자재를 생산하는 중소기업을 대상으로 특허 문서를 수집하고 이를 대상으로 LDA 모델링을 적용해 조선산업 기술테마를 추론한다. 특허기술에 내재된 요소 기술을 바탕으로 기술테마를 정하게 되면 여기에 집약된 실현 가능한 제품군을 살펴볼 수 있게 된다. 시기 별로 개발된 기술 변화의 추이를 분석할 수 있으며, 나아가 국내 조선산업에 필요한 기술군에 대한 준비 과정까지 살펴볼 수 있다.

고병열, 노현숙(2005)은 특허 등 기술 정보를 산업 분석과 연결하는 시도는 1990년대부터 이미 진행돼 왔다고 설명했다. IPC 기반의 특허 정보가 산업의 메가 트렌드를 분석하는 데 적합하지 않으므로, 다양한 확률 모델을 적용해 기술과 산업 간 연계 구조를 파악하는 연구가 활발히 진행됐다.

기술테마는 중소벤처기업부와 중소기업기술정보진흥원이 발표한 “중소기업 전략기술로드맵 2021-2023”을 기준으로 삼는다. IMO의 선박 환경 규제에 따라 지난 10년 동안 국내 조선산업의 기술이 어떻게 변화했는지를 살펴보고, 기술 로드맵에서 제시한 선박용 기술과의 연관성을 분석한다. 기술 로드맵과 연결된 친환경 기술 테마를 살펴보고 이를 특허와 연결한다. 연도별 친환경 기술 개발의 추이를 분석한다.

Ⅱ. 관련 연구

1. 국내 조선산업과 IMO 규제

조선산업의 주도권 경쟁 역사는 대형화와 자동화로 압축된다(Eich-Born & Hassink,2005). 수작업 중심의 선박 건조 기술을 보유한 영국은 용접과 블록 기술을 앞세운 일본에 주도권을 빼앗겼으며, 1990년대에 한국은 대형 선박 건조용 도크와 3D 캐드 기술을 설계에 접목해 블록의 대형화 공법으로 산업의 주도권을 거머쥐었다(Lim et al., 2017).

공정에서의 원가절감, 즉 규모의 경제를 효과적으로 실현하는 국가가 산업의 주도권을 확보했다는 점에서 조선산업은 대표적인 중저기술 산업군으로 분류된다. 따라서 국가 차원의 산업 육성 지원책이 산업 경쟁력 확보의 주요 동력원이 될 수 있으며, 이에 따라 2000년대 중반 이후 중국이 조선산업 부문에서 급격히 성장하는 배경이 되었다고 볼 수 있다(양종서, 김창록, 2020).

곽기호(2017)는 한국의 조선산업은 과거에 산업 주도권 확보에 결정적 공헌을 했던 공정 효율성 차원에서의 점진적 혁신이 복합시스템 기술을 요구하는 해양 플랜트 부문에서의 경쟁력 확보에 걸림돌이 되었다는 지적을 제기한된 바 있다.

IMO의 환경 규제는 선박 건조 공정 부문에서 이뤄진 점진적 혁신에 변화를 준 계기가 됐다. IMO는 선박의 온실가스 배출량을 2030년까지 40% 감축(2008년 기준), 2050년까지 50% 감축하는 내용의 규제안을 발표했다(한국해양수산개발원, 2020). IMO 규제안은 단기적으로 선박 에너지의 효율을 개선하는 등의 조처로 온실가스 배출을 오는 2023년까지 30% 이상을 감축하고, 중기적으로 LNG와 수소 혼합연료 및 무탄소 연료 기술을 개발해 2030년까지 40%를 감축해야 한다는 목표안

을 제시했다. 나아가 무탄소 연료 기술 개발이나 연료전지 등 비화석 연료 기술 개발은 물론, 시장기반체제(MBM) 도입으로 2050년까지 50%의 온실가스를 감축해야 한다(해양수산부, 2019).

선박 오폐수에 관한 규제는 이미 오래 전부터 추진됐다(법제처, 2019). 기름과 유해액체물질에 관한 규제는 1980년대에 이미 적용됐다. 선박 오수에 관한 규제는 2003년 효력을 가지기 시작했는데, 이 규제는 점차 강화돼 선박평형수 처리에 관한 규제로 확대됐다. 즉, 특정 해역에서 유입된 선박평형수는 다른 해역에서 배출할 때 미생물이 없는 상태의 정수를 거쳐야 한다. 이 협약은 2004년 IMO에서 채택된 뒤 2017년 발효됐다. 이에 따라 발효일 이후 건조되는 선박은 선박평형수 처리장치를 의무적으로 설치해야 하며, 기존 운항선박은 IOPP(국제기름오염방지증서)를 받은 시기에 따라 선박평형수 처리장치를 설치해야 한다.

<그림 1> IMO 선박 환경 규제 계획안. 한국해양수산개발원 자료 인용.

국내 조선산업 부문에서는 이미 2018년부터 이와 관련한 성과가 나타나고 있다. LNG 추진선 등 관련 선박 수주에서 경쟁국인 중국과 일본을 따돌렸으며, 선박용 기자재에서도 변화의 조짐이 나타났다. 해양 미생물의 해역 간 이동을 규제함에 따라 파생된 기술인 선박평형수 처리장치의 경우 IMO의 최종 승인을 받은 47개 기술 중 국내 기술이 17개를 차지하며 전 세계에서 관련 기술을 가장 많이 보유하게 됐다(해양수산부, 2021).

구분	조건	BWTS 탑재의무일	평형수 처리	
			발효일 이전	발효일 이후
신조선	2017. 9. 8. 이후 건조선박	인도시	D1/D2	D2
현존선	IOPP 갱신 ('14. 9. 8. - '17. 9. 8.)	'17. 9. 8. 이후 첫 IOPP 갱신		
	IOPP 갱신 ('17. 9. 8. - '19. 9. 8.)	'19. 9. 8. 이후 첫 IOPP 갱신		

· D1: Ballast water exchange(선박평형수 교환)
· D2: Ballast water treatment system(선박평형수처리설비)
· IOPP: International Oil Pollution Prevention(국제기름오염방지증서, 유효기간 5년)

<그림 2> IMO의 선박평형수 처리장치 규제안. 법제처(2019) 자료 인용.

따라서, 선박을 건조하는 조선 부문과 선박용 부품을 제조하는 조선기자재 부문의 기술 트렌드의 변화를 함께 살펴볼 필요가 있다. 그동안 특허를 중심으로 한 연구는 대부분 특허의 서지 정보에 기반한 양적 분석에 치우친 한계가 있다. 특정 선박기술에 대한 출원건수와 인용 현황 등 서지 정보를 기반으로 특허 동향을 파악하는 연구가 주로 진행됐다.

김보람, 안영균(2021)은 선박 관련 기술 중 4차 산업혁명과 관련된 키워드를 중

심으로 특허의 출원 건수와 출원인을 비교하고, 특허맵을 작성하는 연구를 수행했다.

Hu 등 (2019)은 특수목적선, 조선기자재, 일반 선박 등으로 카테고리를 나눠 특허 출원 건수와 양적 성장률 등을 분석해 특허 기술 동향을 파악한 바 있다.

조선기자재 부문을 중심으로 한 조선산업은 동남권 지역 산업의 중추적인 역할을 담당한다. 특히 산업 혁신을 위한 인프라 구축 부문에서 경쟁력을 가진 지역으로 평가받는다(부산산업과학혁신원, 2019). 따라서 특허의 서지 정보에 기반한 기존 연구에서 벗어나 친환경 기술을 중심으로 특허가 가진 잠재적 특성을 파악하는 형태의 분석이 필요한 것으로 파악된다.

2. 토픽모델링

토픽 모델링은 대량의 문서를 분석하고 이해하는 데 유익하다(Hu et al., 2013). 문서 집합에서 확률적 방법론을 적용해 유사한 의미를 가진 단어를 모으고, 이를 토대로 토픽을 추론하는 방법론으로 널리 활용된다.

대표적인 토픽 모델링 알고리즘인 LDA(Latent Dirichlet Allocation)의 이론적 출발점은 문서가 잠재적 토픽의 확률적 집합으로 표현되며, 각각의 토픽은 문서 내 단어의 분포에 따라 특성이 결정된다는 가정에서 시작한다(Blei et al., 2003). 문헌은 여러 개의 토픽을 포함할 수 있으며, 토픽에는 여러 개의 단어가 포함될 수 있다. 문서에 존재하는 개별 단어는 특정 주제에 포함된다. 문헌-주제-단어에 대한 세 단계의 계층적 베이즈 확률 이론을 적용해 문헌에 어떤 토픽이 들어갈지를 미리 정한 다음, 하나의 토픽을 골라 이에 포함된 단어를 골라 문헌에 추가하는 방식으로 모델의 계산이 반복된다. 각각의 문서에 포함된 단어가 어떤 토픽과 연결되는지 결정하고 토픽에 포함될 확률을 산출한다. 토픽을 구성하는 단어의 집합을 추출하

는 과정이다. 토픽을 구성하는 단어의 기여 수준이 확률분포로 표현되므로 토픽의 의미를 추론하고 해석하는 데 장점이 있는 모델로 평가받는다.

특허에 내포된 단어의 확률적 정의로 표현되는 LDA의 특성상 특정 산업군에 속한 기술에 대한 숨겨진 의미를 파악할 수 있다. 따라서 이를 활용해 기술 트렌드를 분석하는 연구가 활발히 진행 중이다.

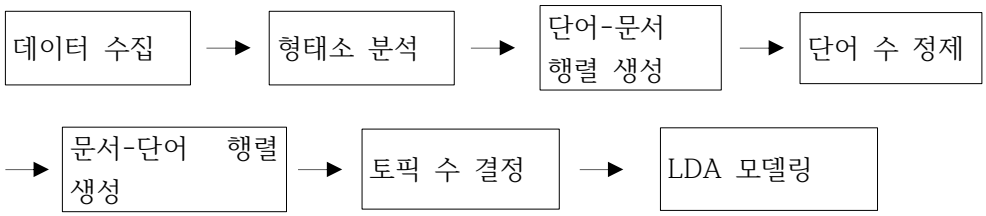
Wang 등(2021)은 중국학술정보원에서 블록체인 관련 특허문서 8,245건을 수집해 LDA 모델링을 적용, 중국의 블록체인 기술이 가진 잠재적 토픽을 추출하고 나아가 최적의 토픽 수를 결정하는 과정을 심도있게 분석했다.

이다혜 등(2018)은 바이오 연료와 관련된 특허 30,000여 건을 분석해 LDA 모델을 통해 바이오 연료 관련 세부 기술을 정의했다. 정의된 세부 기술을 활용해 기술 동향을 파악하고, 세부 기술 간 연결의 정도를 분석했다.

유승태 등(2020)은 국내 산업보안과 관련한 학술논문과 신문기사를 수집해 LDA 모델링을 활용해 기술 트렌드를 분석했다.

이처럼, LDA는 대량의 문서를 기반으로 키워드를 추출하고, 미지의 토픽에 나열된 키워드 집합을 살펴본 뒤 토픽을 추론하는 형태로 분석이 이뤄진다. 토픽을 추론한 뒤에는 토픽과 특허를 연결해 기술 트렌드를 정량적으로 파악할 수 있다.

LDA 모델의 수행 과정은 다음과 같다.



<그림 3> LDA 모델링 순서도

LDA는 단어에 V개의 인덱스({1,...,V})를 부여한 뒤 특정 단어가 V번째에 존재하면 1, 존재하지 않으면 0으로 표현하는 것으로 시작한다. 문서(w)는 N개 단어의 배열로 표현된다. 말뭉치는 M개 문서의 집합으로, $D = \{w_1, w_2, \ldots, w_M\}$으로 표기한다. N은 포아송 분포를 따르며, θ는 디리클레 분포를 따른다. 문서 내 단어에 대해 토픽 z_n는 multinomial(θ)를 선택하며, z_n이 주어지면 w_n은 $p(w_n|z_{n,\beta})$로부터 구한다.

D.M.Blei(2003)가 제시한 기본 가정은 다음과 같다.

1. Choose N ~ Poisson(ζ).
2. Choose θ ~ Dir(α)
3. 각각의 N 단어가 문서에 존재할 때,
   a) Choose a topic z_n ~ Mutinomial(θ)
   b) Choose a w_n from $p(w_n|z_{n,\beta})$

여기서, α는 k차원 디리클레분포의 매개변수다. θ는 k개의 벡터로, 문서가 특정 주제에 속할 확률 분포를 뜻한다. 마찬가지로 k개의 벡터를 가진 z_n은 단어 w_n이 특정 주제에 속할 확률 분포를 나타낸다. β는 $k \times V$ 크기의 행렬 변수로, β_{ij}는 I번째 주제가 j번째 단어를 생성할 확률을 뜻한다.

먼저, 디리클레 변수 θ의 확률밀도함수는 다음의 식을 따른다.

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_1^{\alpha_{k-1}}\cdots\theta_k^{\alpha_k-1}, \quad (1)$$

파라미터인 α와 β가 주어졌을 때의 θ, z, w의 결합 확률 분포는 다음과 같이 구해진다.

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha)\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta), \quad (2)$$

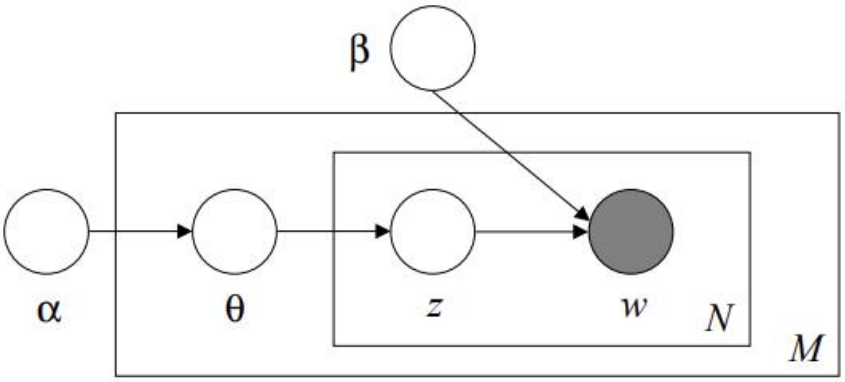
z_n과 θ를 모두 더해 주변 확률(marginal probabilities)을 구할 수 있다. 주변 확률은 결합 사건의 합으로 표시되는 확률이다.

$$p(w|\alpha, \beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\theta)p(w_n|z_n, \beta)\right)d\theta, \quad (3)$$

마지막으로 개별 문서에 대한 주변 확률을 모두 곱해 말뭉치의 확률을 구할 수 있다.

$$p(D|\alpha, \beta) = \prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn}, \beta)\right)d\theta_d, \quad (4)$$

α와 β는 말뭉치 단위의 파라미터로 표현된다. θ_d는 문서 단위의 변수로, z_{dn}과 w_{dn}은 단워 단위의 변수로 표현된다. 이를 도식화하면 아래와 같다.



<그림 4> LDA 모델링 과정. M.D.Blei(2003)

Ⅲ. 사례 분석

1. 가설의 설정과 데이터 수집

선행 연구를 통해 다음과 같은 가설을 세운다.

가설 1. IMO의 환경 규제가 친환경 기술 개발 증가에 영향을 미쳤을 것이다.

가설 2. 친환경 기술은 비(非)친환경 기술보다 더욱 활발한 개발이 이뤄졌을 것이다.

본 연구에서는 2009~2019년 동안의 선박 관련 특허 13,673건을 수집해 LDA 기반의 토픽 모델링을 수행한다. 수집 대상은 한국조선기자재공업협동조합 회원사 255개사다. 한국조선기자재공업협동조합은 전국 단위의 선박 관련 부품을 납품하는 기업이 모인 곳으로, 조선사를 포함해 전통적인 선박 관련 부품 제조사와 IT 관련 기업까지 포함하고 있어 조선산업과 관련한 특허 트렌드를 분석하기에 가장 적합한 것으로 판단했다.

먼저 한국조선기자재공업협동조합의 회원사 리스트를 확보한 뒤, 특허정보검색 홈페이지(KIPRIS)에서 기업명과 출원인을 비교했다. 이를 바탕으로 특허 검색 기간을 2009~2019년으로 설정했다. 실용신안과 디자인, 상표는 검색에서 제외했다. IMO의 선박 환경 규제가 공개 이후의 시기로, 친환경 선박 기술과 관련한 특허를 분석하기에 적합하다. 특허는 등록, 취하, 출원을 모두 포함했다. 친환경 기술 개발의 틀 안에서 특허의 포괄적 특성을 이해하기 위해 취하된 특허도 분석 대상으로 삼았다.

2. 형태소 분석

곽수정 등(2013)은 자연어 처리 과정에서 형태소 분석은 가장 기본적인 단계로, 특히 한국어는 다양한 음운 현상이 생겨 분석 과정이 매우 복잡한 것으로 평가받는다고 설명했다. 따라서 기분석 사전을 이용하면 별도의 연산 과정 없이 효율적으로 형태소 분석을 실행할 수 있다.

하지만 형태소 분석에서 불용어 처리의 기준은 모호하다. 불용어 리스트는 알고리즘 분석에서 파생되는 노이즈를 제거하는 데에는 효과적이지만, 불용어 자체가 안고 있는 연구자의 주관성과 연구 결과의 편향성 문제는 회피하기 어렵다 (Blanchand & Antonie, 2007).

박민규 등(2021)은 LDA 모델링을 활용한 핀테크 관련 특허를 분석한 연구에서 기술 관련 용어를 정리하는 데 다음과 같은 세 가지 문제점을 지적했다. 첫째, 핀테크 용어를 금융 산업에 한정지어 적용하면 연구의 관점과 의견이 달라 단편적인 정의가 이뤄질 수 있다. 둘째, 핀테크 기술 수집 관점에서 IPC 분석 전문가에 의한 분석이 체계적으로 진행되지 않았다는 한계가 있다. 셋째, 핀테크 기술의 적용 관점에서 기술 자체에 집중해 분석이 이뤄졌으므로 핀테크 기술의 산업 파급 효과에 대한 심도 있는 분석이 부족했다.

조선산업 역시 최근의 환경 규제로 말미암아 새로운 기술이 등장하고 있어, 제대로 된 용어가 정립되지 않았다. 선박평형수 처리장치에 관한 용어는 발라스트나 BWMS 또는 BWTS 등으로 혼재돼 있다. 이외에도 액화천연가스와 LNG 등의 용어가 특허 문서에 동시에 등장한다. 특히, 3D 캐드 등의 설계 관련 기술은 물론 자재 관리 시스템이나 모니터링, 자율운항선박 등 IT 관련 기술 융합까지 이뤄지고 있어 기술 자체를 조선산업으로 한정지어 연구를 진행할 경우 연구자의 주관이 강하게 개입될 여지가 있다.

본 연구에서는 기술적 의미를 가진 단어를 최대한 포함하는 방식으로 분석을 진

행한다. 친환경 기술은 선박 관련 기술 전체와 비교할 때 의미를 가질 것이라는 판단에서다.

분석은 R을 활용했으며, 형태소 분석은 NLP4kec를 사용했다. NLP4kec는 명사를 추출하지 않고, 형용사와 동사 형태의 단어도 추출하는 특징을 가진다.

불용어는 빈도수가 높은 단어를 중심으로 기술적 의미를 포함하지 않고 특허 문서 내 문장을 완성하는 의미를 가진 단어를 제거했다. 제거한 단어는 다음 표와 같다.

<표 1> 제거 단어 목록

구분	제거 단어
1	관하다
2	따르다
3	위하다
4	의하다
5	통하다
6	대하다
7	이루어지다
8	부르다
9	메이다
10	이루다
11	가지다
12	용이하다
13	상세하다
14	인하다
15	이러하다

텍스트를 전처리하는 데 있어서 잦은 빈도로 등장하는 단어의 처리에 관한 기준은 모호하다. 높은 빈도로 등장하는 말뭉치가 결과에 좋지 않은 영향을 미칠 수 있어 제거하는 방안도 있지만,[3] 선박 기술과 관련한 전체적인 내용을 이해하는 데에

는 자칫 친환경 관련 기술로 치우친 결과물이 나올 가능성도 있다.

<표 2> 상위 20개 단어 빈도수

구분	단어	빈도수
1	시스템	7128
2	케이블	3866
3	배기가스	2680
4	구조물	2603
5	파이프	2335
6	프레임	2315
7	가이드	2049
8	실린더	2045
9	하우징	1778
10	밸러스트	1728
11	데이터	1648
12	플레이트	1578
13	압축기	1187
14	케이스	1153
15	반응기	1113
16	발전기	1093
17	프로펠러	1007
18	와이어	999
19	베이스	967
20	교환기	898

3) 이소정, “잠재디리클레할당 분석을 이용한 ‘노인일자리’ 관련 신문기사 토픽분석”, 『Journal of Digital Convergence』, Vol.18 No.10(2020), pp.537-546.

가장 잦은 빈도로 등장하는 단어인 '시스템'은 배기가스 처리 시스템일 수도, 선박평형수 처리장치일 가능성도 있다. 데이터와 관련한 기술을 가리키거나, 해양플랜트와 관련된 기술에도 등장하는 단어다. 파이프나 프레임 역시 내연기관 기반의 선박 기술에 자주 등장하지만 친환경 관련 기술에도 적용될 가능성이 있는 단어다. 선박 기술과 관련해 전문적인 지식이 없는 한 기술적 의미를 포함하는 단어를 제거하는 것은 연구자의 주관이 지나치게 많이 개입되는 결과를 초래할 수 있다.

따라서 주관적 개입을 최소화하기 위해 텍스트 마이닝에 같은 의미를 가진 단어를 하나로 뭉치는 작업을 진행했다. R의 gsub 함수를 활용해 50여 개의 단어를 수정했다. '라스트'나 '러스트'와 같은 단어는 전문가의 조언에 따라 선박 기술에 존재하지 않는 부품이라는 것을 확인하고 '밸러스트'에 포함시켰다. 이외에도 지게차와 굴삭기는 중장비로 통합했고, 같은 의미인 엘엔지와 액화천연가스를 같은 단어로 처리했다.

2000여 개의 단어를 불용어로 정의하고 제거를 한 결과물과 50여 개의 비슷한 의미를 가진 단어를 처리한 결과물을 비교해봤다. LDA 모델링을 거쳐 특정 테마에 모인 키워드가 얼마나 일관성 있게 모였는지를 분석했다. 토픽의 개수를 정의하는 k값은 같은 조건을 적용했다. 분석 결과의 일부는 다음과 같다.

<표 3> 불용어 2000여 개를 제거한 토픽-키워드 결과물

	키워드1	키워드2	키워드3	키워드4	키워드5
토픽1	시스템	배터리	에너지	발전기	인버터
토픽2	파이프	플랜지	플레이트	프레임	케이블
토픽3	시스템	데이터	모니터링	카메라	실시간
토픽4	베어링	프로펠러	샤프트	감속기	구동축
토픽5	배기가스	시스템	연료전지	연료유	터보차저
토픽6	시스템	교환기	기화기	액화천연가스	천연가스
토픽7	조성물	케이블	우수하다	아크릴	에틸렌
토픽8	케이블	광섬유	감싸다	둘러싸다	광케이블
토픽9	발전기	구조물	블레이드	시스템	에너지
토픽10	케이블	트레이	서포트	모델링	구조물

<표 4> 동의어 처리 결과물

	키워드1	키워드2	키워드3	키워드4	키워드5
토픽1	시스템	데이터	모델링	시뮬레이션	데이터베이스
토픽2	크레인	프레임	와이어	시스템	리프팅
토픽3	구조물	케이스	파이프	플로터	중량물
토픽4	밸러스트	시스템	중화제	샘플링	미생물
토픽5	파이프	라이저	드리다	크레인	시추선
토픽6	케이블	트레이	서포트	구조물	모델링
토픽7	가이드	캐리지	와이어	파이프	용접부
토픽8	배기가스	시스템	세정액	황산화물	전처리
토픽9	케이블	조성물	우수하다	친환경	할로겐
토픽10	케이블	광섬유	감싸다	둘러싸다	절연체

특정 토픽을 중심으로 모인 키워드에 관해 일관성을 중심으로 살펴본 결과 두 집단 사이에 특별한 차이점이 없었다. 연구자의 주관성 개입이라는 측면에서 불용어의 개념을 적극적으로 정의하지 않더라도 유의미한 결과를 도출할 수 있다는 점을 발견했다.

3. 텍스트 정제 과정

기술테마를 만들기 위한 LDA 모델링 과정은 크게 형태소 분석 → 말뭉치(corpus) 구축 → 단어-문서 매트릭스 생성 → 문서-단어 매트릭스 생성 → LDA 분석으로 나뉜다.

말뭉치 구축은 특허문서 내에 포함된 요약문을 활용했다. 발명의 내용은 지나치게 방대한 양의 정보를 포함해 기술 특성을 효과적으로 추출하기 어려운 단점이 있는 반면, 요약문은 기술 속성에 관한 핵심적인 내용을 담고 있어 기술테마 추출에 효과적이다(최성철, 서원철, 2019).

문서별 핵심 단어를 정제하고 추출해 코퍼스를 구축한 뒤에는 등장 빈도수 중심으로 정렬된 단어-문서 매트릭스를 만든다. 단어-문서 매트릭스는 한 문서에 어떤 단어가 포함되는지를 벡터 형태로 표현한다. 따라서 매트릭스가 매우 큰 차원을 가지게 돼 희소성(sparse) 문제가 중요한 이슈로 부각된다. 이른바 차원의 저주(Curse of Dimensionality)가 존재해 텍스트 데이터의 분석을 어렵게 만드는 원인이 된다(한국은행, 2019).

따라서 출현 빈도수가 매우 작은 단어도 매트릭스에 포함될 수 있으며, 지나치게 많은 단어가 포함될 경우 차원이 기하급수적으로 확대돼 연구 작업의 효율성을 저하시키는 결과가 나타날 수 있다. 특히 특허 기술에 내재된 기술적 함의가 흐려질 수 있다. 이 연구의 텍스트 정제 결과 13,673개의 문서에서 5,649개의 단어가 추출

됐다. 따라서 5,649개 단어 전체를 분석 대상으로 삼는 대신 정제를 통해 기술적 의미를 가진 토픽의 출현 가능성을 높이는 작업이 필요하다.

출현 빈도에 따라 단어를 정렬한 뒤 누적 비율을 산출했다. 그 결과 680개의 단어가 누적비율 80%에 포함됐으며, 1,360개의 단어가 90%를 차지했다. 본 연구에서는 단어의 최소 출현 빈도 기준을 누적비율 85% 수준으로 정한 뒤, 출현 빈도 상위 950개 단어를 분석 대상으로 삼았다.

950개 단어의 출현 빈도수를 기반으로 문서-단어 매트릭스를 만들어 LDA 분석을 위한 준비 작업을 마쳤다.

4. LDA 모델링

Cvitanic 등(2016)은 LDA 모델에서 토픽의 수 k는 사전에 결정하는 파라미터라고 설명했다. k값은 토픽 모델의 품질에 영향을 미친다. 토픽의 개수를 산출하기 위해 일반적으로 활용하는 것으로 Perplexty가 꼽힌다. Perplexty는 토픽의 수가 늘어나면 감소하는 경향을 보이지만, Perplexty가 낮다고 해서 토픽의 일관성이 있다고 해석을 할 수는 없다.

따라서 Perplexity 외에도 k값을 산출하기 위한 다양한 이론이 제시됐다. 마르코프 체인 몬테카를로 알고리즘을 적용하거나, 단어 간 유사함의 정도를 측정해 토픽의 해석을 용이하게 하는 Coherence Score를 제시한 연구도 있다(David, N. et al, 2010).

본 연구에서는 Griffifths(2004), Arun(2010), Cao(2009), Deveaud(2014)의 이론을 적용한 R의 FindTopicsNumber 함수를 사용했다.

1) Griffifths2004

함수에 적용된 Griffifths2004 지표는 마르코프 체인 몬테카를로 알고리즘을 적용했다. *T*개의 토픽이 주어졌을 때, 주어진 문서 속의 i번째 단어의 확률은 다음 식으로 구할 수 있다.

$$P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j)P(z_i = j), \quad (1)$$

z_i는 i번째 단어가 포함된 토픽을 나타내는 잠재적 변수다. $P(w_i|z_i = j)$는 j번째 토픽에서의 단어 w_i의 확률이다. $P(z_i = j)$는 j 토픽에 포함되는 단어를 선택할 확률이다.

Griffifth2004의 특징은 ϕ와 θ를 파라미터로 잡아 값을 추정하고, 단어가 토픽에 포함될 사후확률 $P(z|w)$을 추정하는 것이다. $P(z|w)$을 산출하기 위해 몬테 카를로 알고리즘을 활용한다.

ϕ의 확률은 다음 식으로 표현된다.

$$P(w|z) = \left(\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\right)^T \prod_{j=1}^{T} \frac{\prod_W \Gamma(n_j^{(W)} + \beta)}{\Gamma(n_j + W\beta)}, \quad (2)$$

θ의 확률은 다음 식으로 표현된다.

$$P(z) = \left(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\right)^D \prod_{d=1}^{D} \frac{\prod_j \Gamma(n_j^{(d)} + \alpha)}{\Gamma(n^{(d)} + T\alpha)}, \quad (3)$$

(2)에서 $n_j^{(w)}$는 단어 w가 토픽 j에 할당되는 횟수를 의미한다. 식(3)에서의 $n_j^{(d)}$는

문서 d에 포함된 단어가 토픽 j에 할당되는 횟수다. 사후확률을 구하기 위해 다음과 같은 식을 도출한다.

$$P(z|w) = \frac{P(w,z)}{\sum_z P(w,z)}, \tag{4}$$

식(4)는 분모의 합을 직접적으로 구할 수 없으므로, 마르코프 체인 몬테 카를로 알고리즘을 적용한다.

$$P(z_i = j|z_{-1}, w) \propto \frac{n^{(w_i)}_{-i,j} + \beta}{n^{(\cdot)}_{-i,j} + W\beta} \frac{n^{(d_i)}_{-i,j} + \alpha}{n^{(d_i)}_{-i,\cdot} + T\alpha}, \tag{5}$$

식(5)의 첫 번째 분수는 토픽 j에서의 w_i의 등장 확률을 나타낸다. 두 번째 분수는 문서 d_i에서의 토픽 j의 확률을 구한다. z로부터 ϕ와 θ를 측정할 수 있다.

$$\hat{\phi}^{(w)}_j = \frac{n^{(w)}_j + \beta}{n^{(\cdot)}_j + W\beta}, \tag{6}$$

$$\hat{\theta}^{(d)}_j = \frac{n^{(d)}_j + \alpha}{n^{(d)}_\cdot + T\alpha}, \tag{7}$$

여기서 측정된 값으로 새로운 w와 z의 분포를 예측할 수 있다.

2)Arun2010

Arun2010의 특징은 LDA를 행렬 인수분해 메커니즘으로 본다는 사실이다. 말뭉치는 두 개의 행렬 인수 M1과 M2로 분할한다. 문서를 D, 토픽을 T, 단어를 W라고 할 때, 문서-단어 빈도 행렬 M은 T*W 차수의 토픽-단어 행렬 M1과 D*T 차수의 문서-토픽 행렬 M2로 나타낸다. 분리된 두 행렬 모두 확률적 의미를 나타낸다. M1은 주제의 단어 분포를, M2는 문서의 주제 분포를 각각 표현한다.

$$\sum_{v=1}^{W} M1(t,v) = \sum_{d=1}^{D} M2(d,t) \forall t = 1, \quad (1)$$

각 주제에 할당된 단어 수를 뜻하는 것으로, 단어의 행 합계와 문서에 대한 열 합계를 의미한다.

SVD(Singular Value Decomposition)를 활용해 행렬 M을 세 개의 행렬로 나눈다.

$$M = U * \Sigma * V', \quad (2)$$

V'는 V의 전치행렬이며, M의 고유 벡터를 포함한다. U는 $M'*M$의 고유 벡터를 가진다. 행렬 Σ는 대각 행렬이고, 대각선 항목을 특이값으로 정의한다. 특이값은 벡터를 둘러싸는 타원체의 반축 길이를 의미하는데, 특이값의 분포가 주제의 분산 분포를 나타낸다.

V의 (i, j) 번째 항목(i, j = 1 ~ m)은 다음과 같이 주어진다.

$$(V)_{ij} = (M1)_{ij}/\sigma_i, \quad (3)$$

쿨백-라이블러 발산(Kullback-Leibler Divergence)을 활용해 주어진 조건인 말뭉치(*C*)와 토픽(*T*)으로, LDA의 결과치인 마르코프 행렬 *M*1, *M*2를 측정한다.

$$Measure(M1, M2) = KL(C_{M1} || C_{M2}) + KL(C_{M2} || C_{M1}), \quad (3)$$

CM1은 토픽-단어 행렬인 M1의 특이값 분포다. CM2는 벡터 L * M2를 정규화해서 나온 분포다. L은 말뭉치 내 각 문서 길이에 대한 벡터이며, M2는 문서-토픽 행렬이다.

3) CaoJuan2009

CaoJuan2009는 토픽 간의 밀집도(Dense)를 분석한다. Standard Cosine Distance를 활용해 토픽(*T*) 간 상관분석을 측정한다.

$$corre(T_i, T_j) = \frac{\sum_{v=0}^{V} T_{iv} T_{jv}}{\sqrt{\sum_{v=0}^{V} (T_{iv})^2} \sqrt{\sum_{v=0}^{V} (T_{jv})^2}}, \quad (2)$$

수치가 낮을수록 각 토픽은 독립적이다. Average Cosine Distance를 활용해 모든 토픽의 안정성을 측정한다.

$$ave\,dis(structure) = \frac{\sum_{j=i+1}^{K} \sum_{j=i+1}^{K} corre(T_i, T_j)}{K \times (k-1)/2}, \quad (3)$$

마찬가지로, 낮은 수치를 보일수록 토픽의 구조는 안정적이라고 볼 수 있다.

4) Deveaud2014

Deaveaud2014는 토픽 개수의 선정 등 매개변수가 필요한 LDA의 단점을 극복하기 위해 고안됐다. 잠재적 개념(Latent Concept)이 도입된 모델이다. 단어의 분포를 기반으로 잠재 개념의 수를 자동으로 추정하는 방법이다. 주어진 함수에 대해 n개의 가장 큰 값을 가진 상위 인수를 생성하는 연산자가 정의된다. 토픽 k에서 가장 높은 확률을 가지는 단어 집합 W_k를 얻는다.

$$\hat{K} = \arg max_k \frac{1}{K(K-1)} \sum_{(k,k') \in T_k} D(K||K'), \tag{1}$$

K는 LDA의 주어진 주제의 수이고, T_k는 LDA에서 모델링한 K개의 토픽 집합니다. 쿨백-라이블러 발산으로 두 확률 분포값을 측정한다.

$$D(k||k') = \frac{1}{2} \sum_{w \in W_k \cap W_{k'}} P_{TM}(w|k) \log \frac{P_{TM}(w|k)}{P_{TM}(w|k')} + \frac{1}{2} \sum_{w \in W_k \cap W_{k'}} P_{TM}(w|k') \log \frac{P_{TM}(w|k')}{P_{TM}(w|k)} \tag{2}$$

$T_{\hat{K}}$의 주제 세트는 잠재적 개념 세트로 간주된다. 전체적인 개념의 유사도는 개념어만 고려하여 계산이 가능하므로, 다양한 개념을 단어 묶음으로 취급해 문서 빈도 기반의 측정식이 제안됐다.

$$\sim (T_{\hat{K},m}, T_{\hat{K},n}) = \sum_{k \in T_{\hat{K},m}} \sum_{k \in T_{\hat{K},n}} \frac{|k \cap k'|}{|k|} \Sigma \log \frac{N}{df_w} \tag{3}$$

$|k_i \cap k_j|$는 두 개념이 공통으로 가지고 있는 단어의 수다. df_w는 w의 문서 내 출

현 빈도이며, N은 수집된 문서의 수를 의미한다. 이를 통해 상위의 M 피드백 문서를 생성한다.

$$M = \arg max \sum \sim (T_{\hat{K},m}, T_{\hat{K},n}) \tag{4}$$

소음을 유발하는, 주제와 관련 없는 개념을 제거하기 위해 k에 점수를 매긴다.

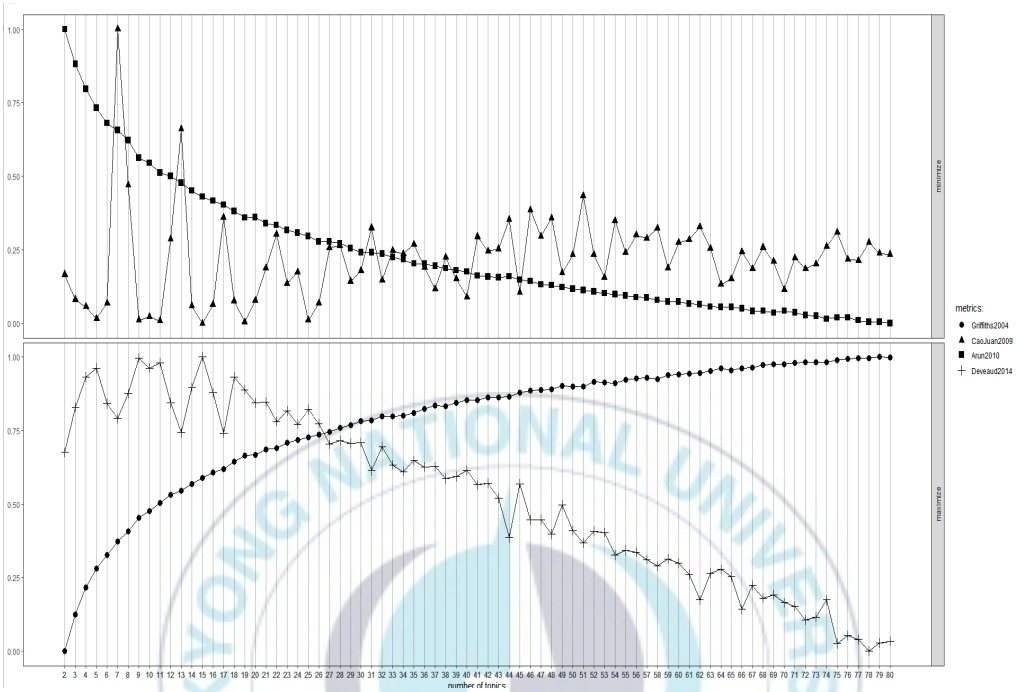
$$\delta_k = \sum_{D \in R_Q} P(Q|D) P_{TM}(k|D) \tag{5}$$

Q는 초기의 쿼리로, 최상위 문서에서 발생하는 개념이 높은 확률로 연관성을 지닌다는 의미다. 문서 D에 나타난 개념 k의 확률 $P_{TM}(k|D)$은 LDA에서 이전에 학습한 다항 분포 θ로 정의된다. 다항 분포 ϕ_k로부터 산출되는 개념 k에 속할 확률을 사용하면 단어 w의 가중치를 계산할 수 있다.

$$\hat{\phi}_{k,w} = \frac{\phi_{k,w}}{\sum_{w' \in W_k} \phi_{k,w'}} \tag{6}$$

5) FindToticsNumber 함수의 활용

CaoJuan2009 지표와 Arun2010 지표는 낮을수록 효과적이다. Griffifths2004와 Deveaud2014 지표는 높을수록 좋은 모델인 것으로 평가받는다. 이 연구에서는 Griffifths와 Arun 지표가 토픽 개수가 많을수록 안정적으로 증가 또는 감소하는 경향을 보인다. 반면 CaoJuan 지표는 토픽 개수가 늘수록 소폭 우상향 하지만, 변동폭이 갈수록 감소하는 모습을 나타냈다. Deveaud 지표는 토픽 개수 증가에 따라 지표가 줄어들었다.



<그림 5> FindTopicsNumber 함수를 활용한 토픽 개수 산출

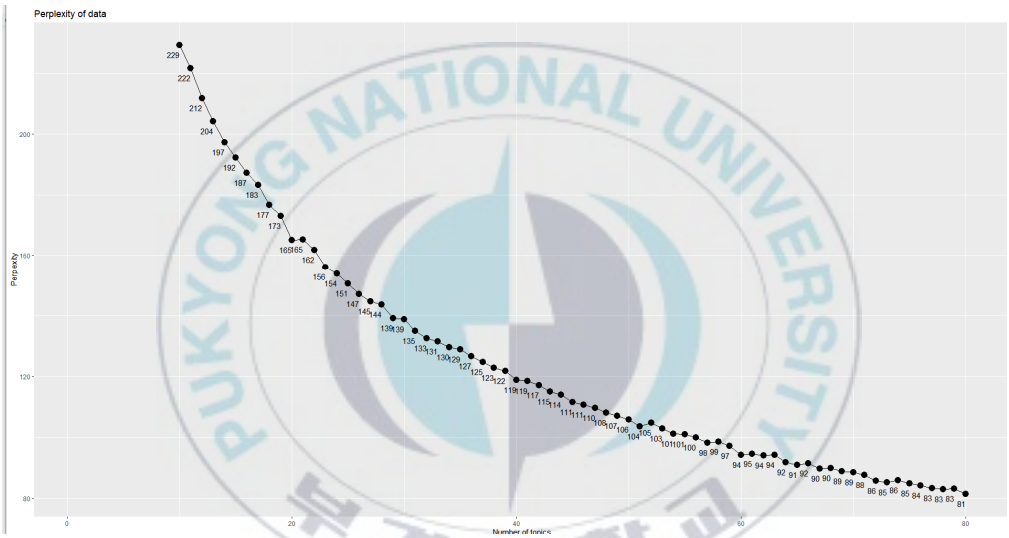
따라서 토픽 개수 15, 24, 45, 64 등 5개의 후보군을 정한 뒤 Perplexity 값을 산출해봤다. Perplexity는 다음의 식으로 측정할 수 있다.

$$Perplexity = \exp\left\{ \frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d} \right\} \quad (1)$$

M은 전체 문서의 수를 의미한다. $p(w_d)$는 문서 d에 포함된 단어의 출현 확률을, N_d는 하나의 문서 d에 포함된 단어의 수를 각각 뜻한다. 즉, 단어가 포함될 확률의 로그합에 대한 전체 문서에 포함된 단어의 개수의 비율에 대한 확률값을 측정하게

되는데 $p(w_d)$가 클수록 Perplexity값은 줄어든다. Perplexity값이 낮을수록 확률의 변동이 낮은 안정적인 모델 결과를 도출할 수 있다.

연구 결과 Perplexity 값은 토픽의 수가 늘수록 안정적으로 우하향하는 모습을 확인할 수 있었다. 토픽 수 15일 때 Perplexity는 197을 나타냈다. 그 값은 급격히 줄어 46에서 111, 64에서 92라는 값을 보였다. 이 과정에 따라 분석할 토픽의 개수를 64개로 정했다.



<그림 6> Perplexity 산출

5. 결과 분석

총 64개의 토픽을 중심으로 분석을 진행했다. 중소벤처기업부가 발표한 『중소기업 전략기술로드맵 2021-2023』를 참고해 조선산업 부문의 8개 친환경 기술을 중심으로 토픽을 추출했다. 그 결과 64개의 토픽 중에서 45개가 친환경 선박 관련 기술인 것으로 집계됐다.

LDA 모델링 과정을 거쳐 토픽에 모인 키워드 20개를 분석했다. 이후 해당 토픽

에 문서에 포함된 특정 키워드가 모인 횟수를 관찰할 수 있는 행렬을 만들어 토픽과 특허를 연결하는 작업을 수행했다.

키워드를 통해 관찰한 결과물을 기반으로 기술테마를 추론하고, 이 추론의 타당성을 검토하기 위해 토픽에 연결된 특허 문서의 수를 산출해 어떤 특허가 특정 토픽에 모였는지를 관찰했다. 그 결과물은 중소기업 기술전략로드맵에서 정한 8개의 친환경 선박 관련 기술을 중심으로 나열했다. 선박 배기가스와 LNG 선박용 기자재 등 친환경 연료 활용 기술에 가장 많은 토픽이 모였다. 연관 기술인 듀얼 추진 선박용 기자재에도 다수의 기술이 집중된 것을 확인할 수 있다.

<표 5> 친환경 선박 기술 테마

기술테마	핵심기술	관련 토픽	높은 관련도를 지닌 키워드(일부)
선박 배기가스 저감장치	·NOx 처리 및 유해가스 제어기술 ·NOx 처리 촉매 기술 ·처리수 제어 기술 ·SOx 흡수율 향상 기술	T8, T26, T30, T47, T52, T58	배기가스, 황산화물, 시스템, 실린더, 열매체, 촉매, 연소, 환원제, 압축기, 반응기
LNG 선박용 기자재	·LNG 화물창 및 운반선 패널 ·LNG 선박용 단열 제품 ·저온용 압력밸브 제어 시스템 ·LNG 연료공급용 전자제어 시스템	T3, T17, T25, T32, T54, T55, T56, T64, T22,	구조물, 파이프, 운반선, 컨테이너선, 가이드, 하우징, 고압가스, 압력가스, 압축기, 천연가스
듀얼 추진 선박용 기자재	·압력제어 ·연료공급 파이프 기술 ·기화가스 연료공급 기술 ·연료분사 밸브 기술	T11, T13, T28, T29, T41, T50, T64	연료전지, 액화천연가스, 베어링, 실린더, 감속기, 샤프트, 컨트롤러, 피스톤, 이산화탄소

	·실린더 헤드 기술		
복합경량 소재 적용 중·소형 선박	·복합소재 적용 선박 하우스 데크 ·CFRP 적용 윙세일 ·알루미늄 다점 곡면가공 기술 ·복합소재 적층구조 선체	T51, T60, T62	조성물, 아크릴, 폴리우레탄, 고분자, 혼합물, 단열재, 보강재, 샌드위치, 극저온, 알루미늄, 와이어
레저선박 기자재 및 레저기기	·선박 구조 설계 ·친환경 추진 기술 ·신소재 선체 소재 기술 ·레저용 안전용품	T19, T39, T40, T49, T53, T59	실리콘, 불순물, 비정질, 발전기, 프로펠러, 추진력, 스러스터, 축전지, 경사지다, 브릿지, 상갑판
오염물 제거 시스템	·선체부착 수중생물 제거 ·수중 코팅 도막 생성 ·선박용오염수 정수처리시설 ·배관 내 잔류생물 부착 처리 기술	T4, T15, T36, T45	밸러스트, 시스템, 중화제, 자외선, 배출관, 컨트롤, 전기분해, 전해조, 정류기
선박충돌 감지 시스템	·충돌감지 및 회피 알고리즘 ·최적항로 결정을 위한 통합 운영 시스템 ·해양환경 계측 모니터링 ·자율운항 선박 운동 예측기술	T32, T35, T37, T44, T61	데이터, 화물창, 가이드, 모델링, 모니터링, 카메라, 이미지, 실시간, 안테나, 주파수, 초음파, 레이더
기관실용 기자재	·고효율 냉각기술 ·다중센서 기반 자동화재 감시 시스템 ·기관실 환기시스템	T23, T24, T27, T31, T35, T38, T48, T50	정화통, 파이프, 케이블, 프로그램, 단말기, 네트워크, 케이블, 다이오드, 웨이퍼, 실린더

※중소기업 전략기술 로드맵 재정리

선박 배기가스 저감장치는 엔전의 연소 과정에서 생기는 배기가스를 줄이는 장치다. 세정기술(T8), 촉매의 공급 및 분사 기술(T47), 환원제의 제어 기술(T58) 등이 포함됐다.

LNG 선박용 기자재는 기존 선박에서 사용했던 화석 연료를 대체해 LNG를 연료로 사용하는 선박에 관한 기술이다. LNG 추진선박, LNG 벙커링 선박과 인프라, LNG 운반선 등이 이 기술에 해당하는 것으로 정의된다. LNG 연료 공급을 위한 제어 시스템(T25), LNG 선박 관련 기자재(T32), 압력제어(T64) 등이 세부 기술군에 속한다.

듀얼추진 선박용 기자재는 천연가스와 디젤 등 서로 다른 연료를 사용해 추진하는 선박에 적용되는 기술이다. 이중연료 기술은 나아가 대체연료 추진이나 수소 연료 추진 방식으로 기술 발전이 기대되는 영역이다. 주요 기술은 하이브리드 추진 시스템(T11), 이중연료 엔진 시스템(T64), 고압용 밸브(T41) 등이 꼽힌다.

복합경량소재 적용 중·소형 선박과 레저선박 기자재 및 레저기기 등 두 개의 기술테마는 국내에서 주를 이루던 대형 선박 건조 기술 체계에서 벗어나 소형 선박에 친환경 기술을 도입하는 영역이다. 선박 구조 설계를 토대로 신기술을 입히는 방식으로, 화학물질 기반의 내구성 강화 기술(T51)이 복합경량소재 적용 중·소형 선박을 대표한다. 신재생에너지를 적용한 추진 및 발전 시스템(T59), 태양전지 관련 기술(T11)도 레저선박 기자재에 속하는 기술군이다.

오염물 제거 시스템에는 선박 평형수 처리장치 기술(T4, T15, T36, T45)이 포함됐다. 자외선 처리 기법 등 미생물 및 해양 오염원을 처리하는 방식이 각각 다른 것으로 나타났다.

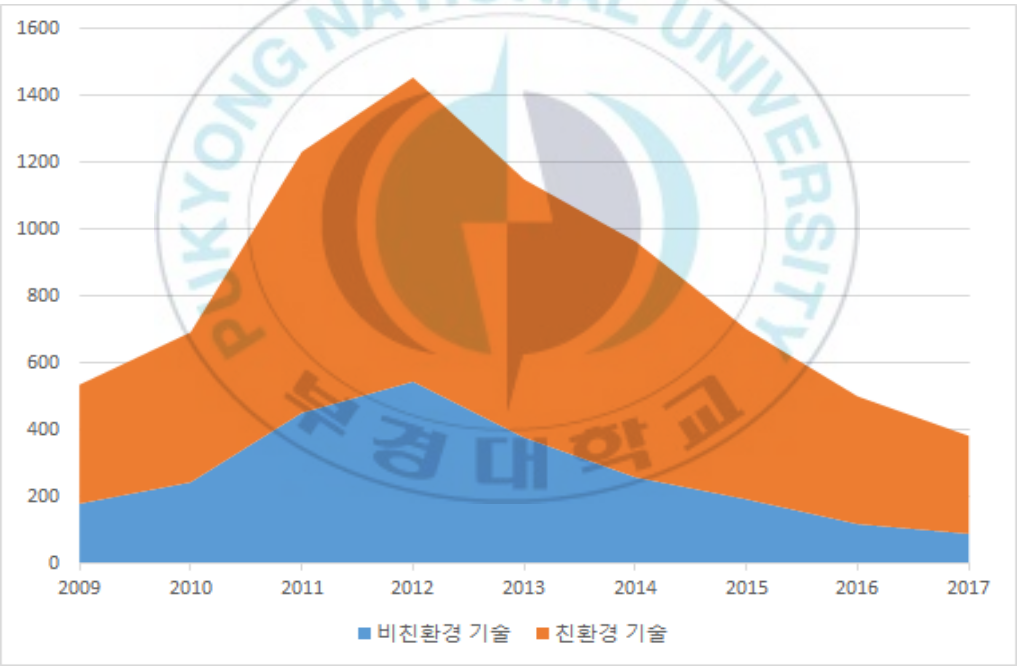
선박 충돌 감지시스템은 자율 운항 선박에 필수적인 기술로 고부가가치 시장에 진출할 가능성이 높다. T35는 해상 상황 모니터링을 의미하고, T37은 선박 운항 제어 관련 기술이 집중됐다.

마지막으로 기관실용 기자재는 기관실에 적용되는 환기시스템, 화재 방지, 배관

점검, 냉각수 점검 등의 기술을 뜻한다. ICT 융합화에 따라 핵심 기관 부품 개발이 요구된다. 선박 배전 감시 장치(T31) 등 선박 운항에 필요한 기관의 모니터링 장비가 다수 포함돼 있다.

Ⅳ. 결 론

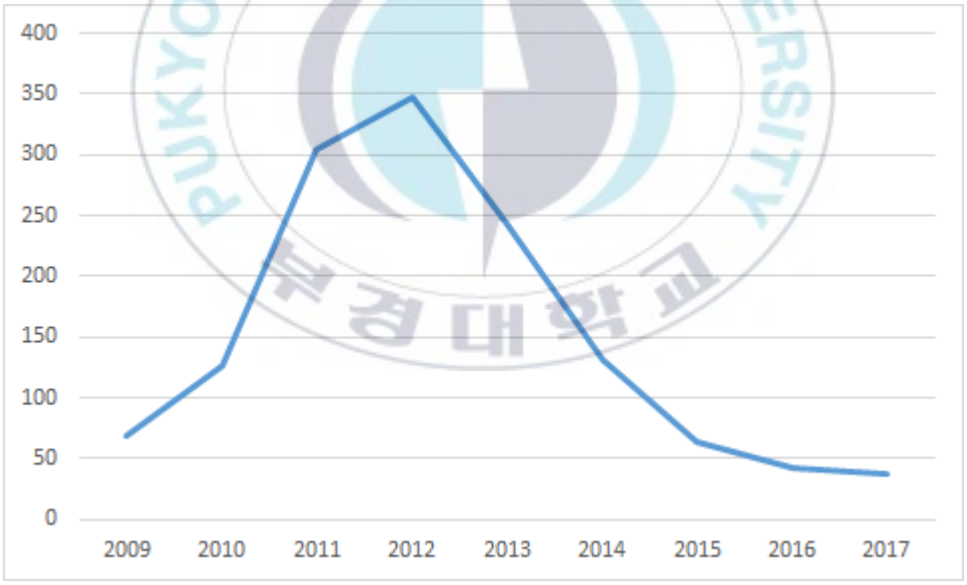
친환경 선박 기술의 변화 양상도 산업 차원에서 관찰할 필요가 있다. 출원일을 기준으로 친환경 기술과 친환경 기술로 분류되지 않은 토픽을 구분한 뒤, 각각의 연도별 성장률을 계산했다. 관찰 시기는 2010~2017년으로 정했다. 이후의 시기는 아직 특허가 등록 또는 공개되지 않았기 때문이다.



<그림 7> 연도별 친환경 선박 기술테마와 나머지 기술 테마의 특허 트렌드 비교

그림에서 보는 바와 같이 친환경 기술 개발이 그렇지 않은 집단보다 압도적으로 많은 것을 확인할 수 있다. IMO의 선박 환경 규제가 2000년대 중반에 발표된 이후

끊임없이 기술 개발이 이뤄진 것으로 추론할 수 있다. 따라서 가설 1, IMO의 환경 규제가 국내 조선산업의 친환경 기술 개발을 유도했을 것이라는 가정을 충족한다. 친환경 관련 특허 수는 2012년 정점에 달했는데, 2009년 대비 두 배 이상 증가했다. 이는 앞서 논의했던 기술테마와 토픽 간의 관계에서도 확인할 수 있다. 전체 64개 중 45개 토픽이 친환경 관련 기술 테마 8개에 집중했다. 이 중, IMO 규제와 직접적으로 연관된 기술 테마인 선박배기가스 저감장치, LNG 선박용 기자재, 오염물 제거시스템에 19개의 토픽이 모였다. 또, IMO 규제로 파생된 기술인 듀얼추진 선박용 기자재와 레저선박 기자재 및 레저기기에 13개의 토픽이 모였다. 모두 합하면 32개의 토픽이 집중된 것으로, 친환경 관련 기술에 포함된 45개의 토픽 중 32개(71.1%)가 IMO 환경 규제와 관련한 것이다.



<그림 8> 연도별 해양플랜트 관련 특허 건수

2012년을 기점으로 두 집단 모두 특허가 감소한 것을 확인할 수 있는데, 이는 해양플랜트 부문에서의 실패가 원인이 된 것으로 추론된다. 이 연구 결과, 친환경 기술로 분류되지 않은 20개의 토픽 중에서 해양 구조물 등 해양플랜트와 관련된 토픽은 11개로 집계됐다. 2012년 350건에 육박했던 해양플랜트 관련 특허는 이후 점차 줄어 2017년 단 38건의 특허만이 이 연구에 등장했다.

2012년 이후 전반적인 기술 개발 활동이 위축되긴 했으나, 친환경 기술은 여전히 그렇지 않은 집단보다 기술 개발 활동 부문에서 상당한 우위를 보인다는 사실을 확인했다. 따라서 친환경 기술이 비(非)친환경 기술보다 더욱 활발한 개발이 이뤄졌을 것이라는, 가설 2에서 제기한 주장도 설득력을 얻는다.

Ⅴ. 논의 사항

이 연구의 의의는 그동안 양적 분석을 중심으로 이뤄졌던 조선산업 특허 분석에서 벗어나, 특허의 내재적 의미를 파악하는 질적 영역으로 확장한 데 있다. 조선산업 부문에서 친환경 관련 기술이 대거 등장한 시기는 2010년부터 2013년 사이로, 이 시기에 배기가스 저감기술, LNG 관련 기술, 선박 평형수 처리장치, 엔진 연료장치 등이 개발된 것으로 파악된다.

반면 연구의 범위를 국내 특허로 한정지은 것과, 해양플랜트에서의 기술 개발이 친환경 기술 개발을 위한 축적의 과정이었는지를 살펴보지 못한 것은 이 연구가 가지는 한계점이다. 선박과 관련한 친환경 기술의 국가 간 비교를 통해 국내 조선산업 기술적 우위를 평가하지 못했다. 향후 USPTO까지 분석의 범위를 넓힐 필요가 있다. 특히, 국내에서 개발한 기술이 세계 각국의 기술이 모이는 USPTO에 등록됐다는 점은 시장성까지 인정받았다는 의미이므로 연구의 필요성은 더욱 높아진다.

해양플랜트 관련 기술은 선박 건조 중심 기술에서 벗어나 복합제품시스템 혁신을 요구하는 영역이다. 특히, 현재 친환경 선박 관련 기술이 수소 연료를 기반으로 한, 온실가스 배출이 전혀 없는 선박 개발로 확대되고 있는 데다, 자율운항선박까지 등장하고 있어 신기술에 관한 수요는 지속적으로 확대될 것으로 예상된다. 향후 연구의 시야를 확대해 친환경 관련 기술의 성격을 정의하고, 해양플랜트 부문에서의 실패 경험이 친환경 기술 개발에 어떤 영향을 미쳤는지 살펴볼 필요가 있을 것으로 파악된다.

참고 문헌

<단행본 및 보고서>

대한조선학회지(2011), 차세대 선박 분야 기술 특허 분석

법제처(2019), 국제해사기구(IMO) 해양환경규제 협약들 관련 법안

부산산업과학혁신원(2019), 부산 조선기자재산업 혁신방안 연구

산업통상자원부(2021), 20년 조선업 수주, 세계 1위

한국은행(2019), 경제 분석을 위한 텍스트 마이닝

한국해양수산개발원(2020), IMO 온실가스 규제 대응 정책방향 연구

해양수산부(2021), 선박평형수 개발업체 및 IMO 승인 현황

해양수산부(2019), '수소선박 안전기준개발사업' 기획연구 보고서

BNK 경제연구원(2021), BNK 경제인사이트

한국수출입은행(2020), 한중일 조선업 경쟁력 분석 및 전남 중형조선산업 지속발전 전략

<국내문헌>

강지호, 김종찬, 이준혁, 박상성, 장동식(2015), "IoT와 Wearables 기술융합을 위한 특허동향분석", 한국지능시스템학회, 25(3), pp.306-311.

곽기호(2017), "복합제품시스템 혁신의 관점에서 바라본 우리나라 해양플랜트 산업의 난관", Korea Business Review, 21(4), pp.169-197.

곽수정, 김보겸, 이재성(2013), "한국어 형태소 분석을 위한 효율적 기분석 사전의 구성 방법", 정보처리학회, 2(12), pp.881-888.

고병열, 노현숙(2005), "기술-산업 연계구조 및 특허 분석을 통한 미래유망 아이템

발굴”, 기술혁신학회지, 8(2), pp.860-885.

김보람, 안영균(2021), “스마트화 대응 우리나라 선박 기술 실태 조사 연구 : 특허 분석 및 감성 분석을 토대로”, 무역학회지, 46(4), pp.1-18면.

박민규, 전병민, 김종우, 금영정(2021), “용어 확장을 통한 핀테크 기술 적용가능 산업의 탐색 : 네트워크 분석 및 토픽 모델링 접근”, 한국전자거래학회지, 26(1), pp.1-28.

서원철(2021), “기술테마 분석을 통한 기술기회발굴 연구-3D 프린팅 기술 사례를 중심으로”, 지식재산연구, 16(2), pp.205-248.

양해성, 곽기호, 전유수(2021), “한국 조선 산업의 주도권 회복에 대한 탐색적 연구 : 최근 고부가가치 선종 수주 성과를 중심으로”, 한국혁신학회지, 16(1), pp.212-158.

유승태, 박경선, 이윤서, 황성진, 김강석(2020), “LDA 토픽 모델링을 이용한 국내 산업보안 동향 분석”, 한국산업보안연구, 10(2), pp.79-103.

이다혜 , 최하영, 정병기, 윤장혁(2018), “특허 텍스트마이닝을 활용한 바이오 연료 기술 모니터링”, 지식재산연구, 13(1), pp.285-312.

이소정(2020), “잠재디리클레할당 분석을 이용한 ‘노인일자리’ 관련 신문기사 토픽 분석”, 디지털융복합연구, 18(10), pp.537-546.

이지호, 정병기, 고남욱, 오승현, 윤장혁(2018), “특허의 Problem-Solution 텍스트 마이닝을 활용한 기술경쟁정보 분석 방법”, 지식재산연구, 13(3), pp.172-204.

전성해(2011), “앙상블모형을 이용한 공백기술예측”, 한국지능시스템학회, 21(3), pp.341-346.

조용래, 김의석(2014), “특허 네트워크와 전략지표 분석을 통한 기업 기술융합 전략 연구”, 지식재산연구, 9(4), pp.192-221.

최성철·서원철(2019), “기술평가를 위한 평가 참조정보 생성 방안에 관한 연구 - 기술적 속성 유사성 관점에서”, 지식재산연구, 14(2), pp.193-226.

<학술지(해외)>

Arun, R., V. Suresh., C. E. Veni Madhavan., & M. N. Narasimha Murthy. (2010). On finding the natural number of topics with latent dirichlet allocation: Some Observations, *Advances in Knowledge Discovery and Data Mining*, 391-402.

Blanchard, A. (2007). Understanding and customizing stopword lists for enhanced patent mapping, *World Patent Information*, *4(29)*, 308-316.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation, *Journal of Machine Learning Research*, *3*, 993-1022.

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection, *Neurocomputing, 72(7-9)*, 1775-1781.

Cvitanic, T., Lee, B., Song, H. I. Fu, K., & Rosen, D. (2016). Lda v. lsa: A comparison of two computational text analysis tools for the functional categorization of patents, *International Conference on Case-Based Reasonin.*

Deveaud, R., SunJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval, *Document numerique*), *17*, 61-84.

Griffifth, T. L. & Steyvers, M. (2004). Finding scientific topics, *PNAS*, *101*, 5228-5235.

Hu, J, Wu, J., Sheng, Y, Wu, J., & Lei, H. (2019). Analysis of chinese patent technology in ship and marine engineering equipment industry, *Materials Science and Engineering*, *688*.

Hu, Y., Boyd-Graber, J., Satinoff, B., & Smith, A. (2013). Interactive topic modeling, *Mach Learn*, *95*, 423-469.

Jeong, B. (2017). Competitive intelligence analysis of argumented reality technology using patent information, *sustainability, 497(9)*

Korobkin, D. M., Fomenkov, S. A., & Golovanchikov, A. B. (2018). Method of identification of patent trends based on descriptions of technical functions, *Journal of Physics, 1015(3).*

Lim, Chaisung et al., Change in industrial leadership and catch-up cycles in shipbuilding industry, *Asian Journal of Technology Innovation*(2017), pp.1-16.

Eich-Born, M. & Hassink, R. (2005). On the battle between shipbuilding regions in germany and south korea, *Sage Journals, 37(4)*

David, N., Grieser, J. H. L. K., & Timothy, B. (2010). Automatic evaluation of topic coherence in human language technologies, *Association for Computational Linguistics*, pp. 100-108.

Shin, S. & Seo, W. (2017). Identifying new technology areas based on firm's internal capabilities, *Journal of Administrative and Business Studies, 3(3),* 114-121.

Wang, J. Fan, Y., Zhang, H., & Feng, L. (2021). Technology hotspot tracking: topic discovery and evolution of china's blockchain patents based on a dynamic LDA model, *Symmetry, 13(415).*

Yoon, B. & Park, Y., (2004). A text-mining-based patent network: Analytical tool for high-technology trend, *Journal of High Technology Management Research, 15*, pp.37-50

Zhang, Q., Li, C., & Wu, Y. (2017). Analysis of research and development trend of the battery technology in electric vehicle with the perspective of patent, *Energy Procedia, 105,* pp.4274-4280.

A Study on Green Technology Trend by Analyzing Patent-Based Technology Themes

Keon Tae Min

Department of Management of Technology, The Graduate School,
Pukyong National University

Abstract

The IMO(International Maritime Organization) regulations is changing the technologies of shipbuilding and marine equipment sector. The initial IMO's plan on greenhouse gas emission from ships stipulated that CO2 per transport work by international shipping should be reduced 40% compared to 2008 by 2030. Since the IMO regulations has been effective in market, new technologies have been adopted in the ship, like ballast water management system, scrubber, related to LNG system.

This study is on patent analysis of shipbuilding and marine equipment sector in Korea, using Latent Dirichlet Allocation(LDA) algorithm. Finding 64 topics in this study and 47 topics were specified in 'Green Technology' theme what is based on 'Small Business Strategic Technologies Road map 2021-2023', the Korean government's suggestion.

In 2009-2012, Korean patents related to 'green technology' have been steeply increased, and the patent quantities were three times higher than the other patents. And then the patents in the two technology theme were steeply decreased. It is presumed that fails in the offshore plant in 2008-2013. This study aims to discovering technology changes by analyzing the relationships between technology themes that describe specific technological implications as a set of technology attributes.

부 록

1. LDA 분석을 위한 R 코드

```
library(readxl)

#한글텍스트 전처리 패키지
library(tidymodels)
library(tidytext)
library(NLP4kec)
library(stringr)
library(tm)

#특허 데이터 불러오기
data <- read_excel("pat.xlsx", col_names = T, na = "-")
data = subset(data, select = -c(심사진행상태, 발명의명칭, 출원번호, 출원일자, 출
  원인, IPC분류, CPC분류))
colnames(data) <- c('id', 'summary')

#데이터프레임 저장
data <- as.data.frame(data)
glimpse(data)
sum(is.na(data))

#중복행 및 텍스트 공백, 의미없는 텍스트 제거
data <- unique(data)
data%>%nrow
data <- sapply(data, str_remove_all, '\\s+')##공백 제거
data <- as.data.frame(data)
data_range <- data$summary %>% nchar %>% range
print(data_range)
```

```
data$summary[nchar(x = data$summary) < 14]
head(data)

#동의어 처리 과정
data$summary <- gsub("개스킷", "가스켓", data$summary)
data$summary <- gsub("골격계", "프레임", data$summary)
data$summary <- gsub("그라인더", "그라인딩", data$summary)
data$summary <- gsub("다이알", "다이얼", data$summary)
data$summary <- gsub("드라이빙", "드라이브", data$summary)
data$summary <- gsub("라스트", "밸러스트", data$summary)
data$summary <- gsub("러스트", "밸러스트", data$summary)
data$summary <- gsub("방사능", "방사선", data$summary)
data$summary <- gsub("방사성", "방사선", data$summary)
data$summary <- gsub("발란스", "밸런스", data$summary)
data$summary <- gsub("복수기", "응축기", data$summary)
data$summary <- gsub("복수열", "응축열", data$summary)
data$summary <- gsub("복원성", "복원력", data$summary)
data$summary <- gsub("부스트", "부스터", data$summary)
data$summary <- gsub("브레이커", "브레이크", data$summary)
data$summary <- gsub("브레이킹", "브레이크", data$summary)
data$summary <- gsub("브리지", "브릿지", data$summary)
data$summary <- gsub("비대칭형", "비대칭", data$summary)
data$summary <- gsub("수나사", "숫나사", data$summary)
data$summary <- gsub("스러스트", "스러스터", data$summary)
data$summary <- gsub("스모그", "스모크", data$summary)
data$summary <- gsub("스캐너", "스캐닝", data$summary)
data$summary <- gsub("스태커", "스태킹", data$summary)
data$summary <- gsub("스테인리스", "스테인리스강", data$summary)
data$summary <- gsub("스파이럴", "나선형", data$summary)
data$summary <- gsub("스프레이건", "스프레이", data$summary)
data$summary <- gsub("슬라이더", "슬라이딩", data$summary)
```

```
data$summary <- gsub("슬라이드", "슬라이딩", data$summary)
data$summary <- gsub("시뮬레이터", "시뮬레이션", data$summary)
data$summary <- gsub("시설물", "구조물", data$summary)
data$summary <- gsub("쌍방향", "양방향", data$summary)
data$summary <- gsub("아크릴로니트릴", "아크릴", data$summary)
data$summary <- gsub("아크릴산", "아크릴", data$summary)
data$summary <- gsub("아크릴산", "아크릴", data$summary)
data$summary <- gsub("아크릴아마이드", "아크릴", data$summary)
data$summary <- gsub("아크릴아미드", "아크릴", data$summary)
data$summary <- gsub("암모늄", "암모니아", data$summary)
data$summary <- gsub("에너지원", "에너지", data$summary)
data$summary <- gsub("엘엔지", "액화천연가스", data$summary)
data$summary <- gsub("여과수", "여과기", data$summary)
data$summary <- gsub("자기력", "자기장", data$summary)
data$summary <- gsub("중량체", "중량물", data$summary)
data$summary <- gsub("중심부", "중앙부", data$summary)
data$summary <- gsub("굴삭기", "중장비", data$summary)
data$summary <- gsub("굴착기", "중장비", data$summary)
data$summary <- gsub("지게차", "중장비", data$summary)
data$summary <- gsub("측정치", "측정값", data$summary)
data$summary <- gsub("컨텐츠", "콘텐츠", data$summary)
data$summary <- gsub("케이싱", "케이스", data$summary)
data$summary <- gsub("콘트롤", "컨트롤", data$summary)
data$summary <- gsub("클래딩", "클래드", data$summary)
data$summary <- gsub("환경친화", "친환경", data$summary)


#형태소 분석(NLP4kec)
parsed_data <- r_parser_r(data$summary, language = "ko")
parsed_data[1:10]
```

```
#corpus 작업
data_corp <- VCorpus(VectorSource(parsed_data))
inspect(data_corp[1])
data_corp <- tm_map(data_corp, removePunctuation)
data[1,1]
data[2,1]
data_corp[[1]][[2]]
data_corp[[1]][[1]]


#불용어 삭제
dic <- read.table(file = "C:/Users/golmp/OneDrive/바탕 화면/대학원 강의/석사
  논문_조선기자재기술융합/로데이터/한국어불용어100.txt", sep = "\t",
  fileEncoding = "UTF-8")
dic <- dic[ , 1] %>% as.character()
dic2 <- c("관하다", "따르다", "위하다", "의하다", "통하다", "대하다", "이루어지다", "
  부르다", "메이다", "이루다", "가지다", "용이하다", "상세하다", "인하다", "이러하
  다")
dic_data <- tm::tm_map(data_corp, FUN = stripWhitespace) %>%
  tm::tm_map(FUN = removeNumbers) %>%
  tm::tm_map(FUN = removePunctuation) %>%
  tm::tm_map(FUN = removeWords, words = c(dic, dic2))


#TDM
data_tdm <- TermDocumentMatrix(dic_data)
data_tdm%>%inspect
data_tdm

##TDM 파일 생성
```

```
library(slam)
write.csv(as.array(rollup(data_tdm,2)) , "termdocumentmatrix_f.csv", row.names =
  T)

###단어 빈도수 확인
word.count <- as.array(rollup(data_tdm,2))                    #매트릭스 행별 합계구하
  기, 950번째 누계비 80%
word.order <- order(word.count, decreasing = T)[1:950] #많이 쓰인 단어 순서정
  리(단어번호)
freq.word <- word.order[1:950]                                 #단어번호
row.names(data_tdm[freq.word,])                                 #해당단어번호 단어 확
  인



#DTM
dtm_corp <- as.DocumentTermMatrix(data_tdm[freq.word,])
rowTotals <- apply(dtm_corp , 1, sum) #Find the sum of words in each
  Document
data_dtm    <- dtm_corp[rowTotals> 0, ]
dim(data_dtm)
write.csv(as.array(rollup(data_dtm,2)) , "documenttermmatrix_f.csv", row.names =
  T) ##문서별 단어수 기술통계량 산출



##dtm_corp를 매트릭스로 변환해 열 합계 구하고, wordfreq 만들기
dtm_corp %>% as.matrix() %>% colSums() -> wordFreq
wordFreq[1:10]
wordFreq <- wordFreq[order(wordFreq, decreasing = TRUE)]#내림차순 정렬
head(x=wordFreq, n=20)
head(x=wordFreq, n=70)
```

```
wordDF <- data.frame(word = names(x=wordFreq),
                     freq = wordFreq,
                     row.names = NULL) %>% arrange(desc(x =
  freq))#wordFreq 데이터프레임으로 정리


#######word cloud 그리기
library(wordcloud)
library(RColorBrewer)

blues <- brewer.pal(8, "Blues")[-(1:2)]
wordcloud(wordDF$word, wordDF$freq, max.words=500, colors=blues,
  family="AppleGothic")



####K값 범위 추론
library(ldatuning)
library(topicmodels)

result <- FindTopicsNumber(
  data_dtm,
  topics = seq(from = 2, to = 80, by = 1),
  metrics = c("Griffiths2004", "CaoJuan2009", "Arun2010", "Deveaud2014"),
  method = "Gibbs",
  control = list(seed = 12345),
  mc.cores = 3L,
  verbose = TRUE
)
result
FindTopicsNumber_plot(result)
```

```
write.csv(result, "k_plot_f.csv")


# lda format, perplexity

data_lda <- dtm2ldaformat(data_dtm, omit_empty = F)

# optimal number of topics : perplexity
Sys.time()
set.seed(12345)

topics <- c(10:80)
burnin <- 500
iter <- 1000
keep <- 50

perplexity_df <- data.frame(data=numeric())

for (i in topics){
  fitted <- LDA(data_dtm, k = i, method = "Gibbs",
                control = list(burnin = burnin, iter = iter, keep = keep) )
  perplexity_df[i,1] <- perplexity(fitted, newdata = data_dtm)
}

perplexity_df
Sys.time()

k <- 64

# lda run
library(lda)
```

```
lda.result <- lda.collapsed.gibbs.sampler(data_lda$documents,
                                          K = k,
                                          vocab = data_lda$vocab,
                                          num.iterations = 80000,
                                          burnin = 25000,
                                          alpha = 0.01,
                                          eta = 0.01)

Sys.time()


write.csv(t(lda.result$document_sums), "patent-topic_matrix_f_64.csv", row.names =
  F)

print(lda.result)
lda.result$topics
lda.result$document_expects
attributes(lda.result)
dim(lda.result$topics)
top.topic.words(lda.result$topics)
lda.result$topic_sums
dim(lda.result$documents_sums)

write.csv(t(top.topic.words(lda.result$topics,                    num.words=20)),
  "top-topic-words_f_64.csv", row.names = F)
```