

# 임베딩 모델과 지식맵 분석을 활용한 반도체 기술문서 중요도 분류에 관한 연구

홍기완\*, 장항배\*\*

\*중앙대학교 일반대학원 융합보안학과

\*\*중앙대학교 경영경제대학 산업보안학과

ghd9201@cau.ac.kr, hbchang@cau.ac.kr

## A Study on the Importance Classification of Semiconductor Technical Documents Using Knowledge Graphs and Embedding Models

Giwan Hong\*, Hangbae Chang\*\*

\*Dept. of Security Convergence, Graduate School, Chung-Ang University

\*\*Dept. of Industrial Security, Chung-Ang University

### 요 약

4차산업혁명과 함께 기존 산업구조가 급속하게 변화하고 기술패권주의가 심화되면서, 기술 패권의 승패에 따라 국가의 글로벌 경쟁력이 크게 좌우된다. 세계 주요국들은 기술경쟁력 확보를 위해 기술 혁신과 기술연대의 경쟁을 벌이고 있고, 우리나라도 이러한 동향 속에서 적극적인 R&D 연구 투자와 정책적 지원을 통해 미래 산업 분야의 기술경쟁력 확보를 위해 노력하고 있다. 현재 중국에 의한 기술 탈취나 인력 유출이 발생하고 있고, 이는 산업경쟁력 상실로 이어져 막대한 경제적 피해를 야기할 수 있다. 기술경쟁력을 잃지 않기 위해, 반드시 우리의 산업기술 보호 수단도 마련되어야 한다. 선제적으로 중요한 산업기술을 적절히 식별하여 중요도에 따라 보호수단을 이행하는 것이 산업기술 보호의 시작일 것이다. 이에 따라, 본 논문에서는 지식그래프와 임베딩 모델을 활용하여 우리나라의 핵심 산업분야 중 하나인 반도체 분야의 기술문서를 중요도에 따라 수직적으로 분류할 수 있는 방안에 대해 연구하고자 한다.

### 1. 서론

4차산업혁명과 함께 기존 산업구조가 급속하게 변화하고 기술패권주의가 심화되면서, 기술 패권의 승패에 따라 국가의 글로벌 경쟁력이 크게 좌우된다. 세계 주요국들은 기술경쟁력 확보를 위해 기술 혁신과 기술연대의 경쟁을 벌이고 있고, 우리나라도 이러한 동향 속에서 적극적인 R&D 연구 투자와 정책적 지원을 통해 미래 산업 분야의 기술경쟁력 확보를 위해 노력하고 있다. 현재 중국에 의한 기술 탈취나 인력 유출이 발생하고 있고, 이는 산업경쟁력 상실로 이어져 막대한 경제적 피해를 야기할 수 있다.

기술경쟁력을 잃지 않기 위해, 반드시 우리의 산업기술 보호 수단도 마련되어야 한다. 선제적으로 중요한 산업기술을 적절히 식별하여 중요도에 따라 보호수단을 이행하는 것이 산업기술 보호의 시작일 것이다. 이에 따라, 본 논문에서는 지식그래프와 임베딩 모델을 활용하여 우리나라의 핵심 산업분야 중

하나인 반도체 분야의 기술문서를 중요도에 따라 수직적으로 분류할 수 있는 방안에 대해 연구하고자 한다.

### 2. 선행 연구

#### 2.1 지식맵

지식맵(Knowledge Map)은 논문, 책, 보고서 등의 지식을 수학적, 통계적 방법으로 시각화한 것을 말하며, 분야에 따라, 지식지도, 지식매핑, 과학기술맵, 토픽맵 등 다양한 용어로 활용되고 있다[1]. 지식관리 측면에서 지식관리 및 비즈니스 프로세스를 향상시키기 위해 내부 지식 및 처리, 구조화 목적이나 사용자가 지식을 네비게이션하기 위해 지식을 가시화하는 목적으로 사용된다[2]. 지식맵은 지식에 대한 정보를 표현하고 관계를 표현하여 지식의 탐색이나 추론 등에 도움을 줄 수 있어 다양한 분야에서 활용되고 있다.

## 2.2 임베딩 모델

임베딩은 자연어를 기계가 이해할 수 있도록 0과 1의 수치로 표현하는 방법을 말한다. 통계적 모델인 TDM(Term-Document Matrix), TF-IDF(Term Frequency-Inverse Document Frequency), 단어 수준의 모델로 Word2Vec, FastText, ELMo 등이 있고, 보다 발전된 형태로 문장 수준 모델이나 뉴럴 네트워크 기반의 모델들이 활용되고 있다[3]. 통계적 모델이나 단어 수준의 모델을 이용하여, 단어의 통계적 수치나 유사도, 동시출현 관계 등을 분석하여 단어 간의 관계를 수치화 할 수 있다.

## 3. 연구 방법론

본 논문에서는 그림 1과 같이 임베딩 모델과 지식맵 분석을 활용하여 기술정보 등급모형에 따른 반도체 기술문서 중요도 분류 방안을 설계하였다. 반도체 기술 분야 논문, 연구보고서, 신문 데이터를 수집하고, 통계적 기반 임베딩 모델과 단어수준의 임베딩 모델을 활용하여 기술문서들의 주요 키워드 간의 관계를 수치화한 뒤 지식그래프를 생성한다. 기업이나 조직의 특성을 고려할 수 있는 중요단어사전을 설정한다. 임베딩 모델을 통해 생성된 지식맵을 기반으로 수집된 문서들에서 등장한 중요단어를 분석한다. 등장한 중요단어에 기초하여 문서의 중요도 점수를 산출하고자 한다.

통계적 모델 결과를 이용하여 개별단어 노드의 크기를 표현하고, 단어수준 모델 결과를 이용하여 동시출현 관계, 유사도 등을 반영하여 개별단어 간의 간선을 연결하고자 한다. 또한, 국가핵심기술로 선정된 반도체 기술 분야(10개)에서 중요단어사전을 설정하여 수집된 문서데이터별로 중요단어사전 기반 TDM을 생성한다. 생성된 TDM과 지식그래프를 활

용하여 등장한 단어의 계층점수, 동시출현점수, 등장빈도점수를 합산하여 문서별 중요도를 산출하고자 한다. 산출된 문서 중요도의 분포를 통해 수직적 문서 분류 가능성을 탐색적으로 분석하고자 한다.

## 4. 결론 및 향후 연구

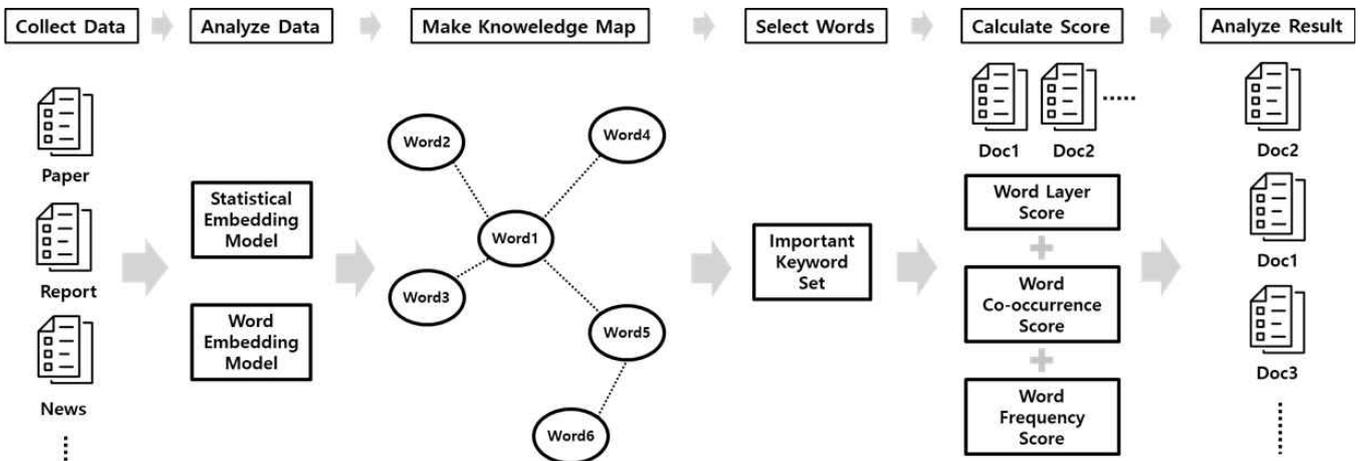
본 논문에서는 기술패권주의 속에서 우리나라의 산업기술 보호 역량 강화를 위해, 기술문서의 중요도 기반 수직적 분류 방안을 설계하였다. 향후 연구로는, 제안한 중요도 기반 분류 방안에서 계층점수, 동시출현점수, 등장빈도점수 산출 방안을 구체화하고, 실험을 수행하여 중요도 분류 가능성에 대해 탐색적 분석을 진행하고자 한다.

### 감사의글

이 논문은 2021년도 정부(산업통상자원부)의 재원으로 한국산업기술진흥원의 지원을 받아 수행된 연구임(P0008703, 2021년 산업혁신인재성장지원사업)

### 참고문헌

- [1] 이방래, 이준영, 김도현, 노경란, 양명석, 권오진, ... &김한준. (2013). 과학기술 지식맵의 형태적 분류와 정보분석 관점의 지식맵 사례 도출. 한국콘텐츠학회논문지, 13(11), 461-476.
- [2] 유기동, &황현석. (2012). 온톨로지를 이용한 프로세스 기반 지식지도 구축. 한국산업정보학회논문지, 17(4), 87-97.
- [3] 유승의, “인공지능과 자연어 처리 기술 동향”, 정보통신기획평가원, 2021



(그림 1) 문서 중요도 기반 수직적 분류 방안