지식재산연구 19권 2호, 2024년 6월, 117-141면
The Journal of Intellectual Property Volume 19, Issue 2, June 2024, Pages 117-141.

Print ISSN 1975-5945
Online ISSN 2733-8487

RESEARCH ARTICLE

Automatic Classification of Non-Patent Literature via Patent-Literature Text Mining

Seongwon Kim[1], Donghee Yoo[2], Suwon Lee[3]

[1]Master's Student, Department of Intellectual Property Convergence, Gyeongsang National University, Republic of Korea
[2]Professor, Department of Management Information Systems, Gyeongsang National University, Republic of Korea
[3]Professor, Department of Computer Science, Gyeongsang National University, Republic of Korea

*Corresponding Author: Suwon Lee (leesuwon@gnu.ac.kr)

ABSTRACT

To file a patent or examine a submitted patent, one must perform a prior-art search that includes both patent and non-patent literature. Unlike patent literature, non-patent literature is not standardized and lacks a unified search system, thus necessitating separate searches for patents and non-patents. This renders the process particularly challenging for the latter. Hence, classification methods used in patent literature are applied to non-patent literature in this study, thus enabling a search system that operates in the same manner as patent-literature searches. The proposal includes the application of machine-learning techniques to recommend or automatically assign patent-classification codes to non-patent literature. For example, a process is reviewed in which international patent classificationcodes are automatically assigned to scholarly papers using machine-learning algorithms. Based on analyzing methods that leverage text-similarity and text-classification algorithms, the automatic classification of non-patent literature through patent-literature text mining is shown to be effective and thus warrants further research. Building a database of non-patent literature coded with patent classifications can result in a more efficient prior-art search process by allowing searches under a unified classification system for both patent and non-patent literatures.

KEYWORDS

Patent Literature, Non-Patent Literature, Text Mining, Automatic Classification, Text Similarity, Text Classification

Check for updates

Open Access

Citation: Kim S et al. 2024. Automatic Classification of Non-Patent Literature via Patent-Literature Text Mining. The Journal of Intellectual Property 19(2), 117-141.

DOI: https://doi.org/10.34122/jip.2024.19.2.6

Received: February 01, 2024
Revised: March 14, 2024
Accepted: May 29, 2024
Published: June 30, 2024

Funding: The author received manuscript fees for this article from Korea Institute of Intellectual Property.

Conflict of interest: No potential conflict of interest relevant to this article was reported.

원저

특허문헌 텍스트 마이닝을 통한 비특허문헌 자동분류에 관한 연구*

김성원[1], 유동희[2], 이수원[3]

[1]경상국립대학교 지식재산융합학과 석사과정, [2]경상국립대학교 경영정보학과 교수, [3]경상국립대학교 컴퓨터공학과 교수

*교신저자: 이수원 (leesuwon@gnu.ac.kr)

차례

국문초록

특허를 출원하거나 출원된 특허에 대한 심사를 수행하려면 특허와 비특허 문헌을 모두 포함하는 선행기술 검색을 수행하는 것이 필수적이다. 특허문헌과 달리 비특허문헌은 표준화되어 있지 않고 통일된 검색 시스템이 없기 때문에 특허와 비특허를 별도로 검색해야 하며, 특히 비특허 문헌의 경우 검색 과정이 까다롭다. 이를 극복하기 위해 특허 문헌에 사용되는 분류 방법을 비특허 문헌에도 적용하여 특허 문헌 검색과 동일한 방식으로 작동하는 검색 시스템을 구축할 것을 제안한다. 이 제안에는 비특허 문헌에 특허 분류 코드를 추천하거나 자동으로 할당하는 머신 러닝 기법의 적용이 포함되어 있다. 한 예로, 머신러닝 알고리즘을 사용하여 학술 논문에 국제특허분류 코드를 자동으로 할당하는 프로세스를 검토했다. 텍스트 유사도 알고리즘과 텍스트 분류 알고리즘을 활용하는 방법을 살펴본 결과, 특허 문헌 텍스트 마이닝을 통한 비특허 문헌의 자동 분류가 효과적이며 추가 연구가 필요하다는 의견이 제시되었다. 특허 분류로 코딩된 비특허 문헌 데이터베이스를 구축하면 특허 문헌과 비특허 문헌 모두에 대해 통일된 분류 체계로 검색할 수 있어 보다 효율적인 선행기술 검색 프로세스가 가능할 것으로 예상된다.

주제어

특허문헌, 비특허문헌, 텍스트 마이닝, 자동분류, 텍스트 유사도, 텍스트 분류

1. 서론

1.1. 필요성과 목적

비특허문헌은 특허를 받을 발명에 대해 이미 공개된 지식을 보여줌으로써 그 신규성을 정당화하기 위해 특허에 인용된 출판물, 기술 표준, 학술 연구 논문, 임상 시험, 서적 등을 말한다.[1] 이는 특허문헌(특허공보 및 공개특허공보)을 제외하고 출원 일자를 소급 인정받을 수 있는 선행문헌이라는 뜻이며, 이외에도 빠르게 변화하는 신기술 분야의 특허 정보 검토 시기의 공백이 발생하는 현상을 해결해 주거나, 특허문헌상에 상세하게 기술되지 않은 고급 정보들을 확인할 수 있다는 여러 가지 이유로 중요성이 높게 인식되고 있다고 한다.[2][3]

이러한 비특허문헌들은, 특허출원을 위해 명세서를 작성할 때 발명의 배경이 되는 기술[4]에 선행문헌 정보를 기재하기 위해서 검색되거나, 특허성 여부를 판단할 때 출원된 발명과 같거나 유사한 종래 기술이 존재하는지를 확인하기 위해서[5][6] 검색된다. 하지만 비특허문헌들은 INID(Internationally Agreed Numbers for the Identification of Data, 서지적 사항의 식별기호) 코드로 표준화되어 검색이 쉬운 특허문헌들[7]과 달리 표준화가 이루어져 있지 않으며, 종류도 다양하므로 검색이 비교적 어렵다. 일반적으로 검색을 위해서 키워드나 색인을 이용하지만, 학술지를 검색할 때는 철자 오류뿐만 아니라 학술지의 기재 방식이 불규칙하므로 발생한 잘못된 구두점이나 콤마, 대소문자와 같은 문법적 일탈이 존재하여 여러 차례의 반복 작업이 필요하다[8].

특허문헌의 경우, 이를 해결하려는 방법으로 CPC(Cooperative Patent Classification, 협력적 특허분류) 및 IPC(International Patent Classification, 국제특허분류) 코드를 부여하여 분류와 검색을 쉽게 하고 있다.[9] 논문의 경우에는 한국연구재단의 연구분야 분류[10]가 있고 특허청에서 국가과학기술표준 분류와 IPC 간의 연계표를 제시해 주고 있지만[11][12], 사용하기에는 두 가지 문제점이 있다. 첫 번째는 연구분야 분류 자체의 오류가 존재한다는 점[13], 두 번째는 논문 검색 서비스 제공자와 학술지, 국가별로 다른 많은 분류 체계가 있다는 점이다. 이외에도 비특허문헌의 경우 종류가 다양하므로, 선행기술조사는 특허문헌 조사와 비특허문헌 조사로

---

* 본 논문은 특허청이 주최한 '제18회 대학(원)생 지식재산 우수논문공모전'에서 최우수상으로 선정된 논문을 수정·보완한 논문입니다.

1) Gema Velayos-Ortega & Rosana López-Carreño, "Non-Patent Literature", *Encyclopedia*, Vol.1 No.1(2021), p. 198.

2) 장경선 외 5인, "해외 지식재산권 데이터 관리현황 조사 및 연구", 특허청, 2006, 113면.

3) 주시형, "특허인용의 등록유지 기간에의 영향 - 한국특허 인용 정보를 활용한 분석", 「지식재산연구」, 제15권 제4호(2020), 309면.

4) 특허법 시행규칙 제21조 제3항.

5) 한국특허기술진흥원, "선행기술조사", 한국특허기술진흥원, <https://www.kipro.or.kr/business/priorArtSearch>, 검색일: 2024. 1. 26.

6) 박영규, "선택발명의 신규성·진보성 판단을 위한 선행기술의 인정범위", 「지식재산연구」, 제14권 제4호(2019), 199면.

7) Wikipedia, "INID", Wikipedia, <https://en.wikipedia.org/wiki/INID>, 검색일: 2024. 1. 26.

8) 권오진 외 4인, "핵심정보자원 연계를 통한 국내 특허 인용 정보 생성 방법에 관한 연구", 한국기술혁신학회 학술대회, 한국기술혁신학회, 2005, 695면.

9) 특허청, "CPC 및 IPC 분류코드", 특허청, <https://www.kipo.go.kr/ko/kpoContentView.do?menuCd=SCD0200269>, 검색일: 2024.01.26.

10) 국가과학기술표준분류, 학술연구분야분류, 전문위원(RB)분야분류가 있다. 한국연구재단, "연구분야분류표", 한국연구재단, <https://www.nrf.re.kr/biz/doc/class/view?menu_no=322>, 검색일: 2024. 1. 26.

11) 특허청, "기술-품목-특허 연계표", 특허청, <https://www.kipo.go.kr/ko/kpoContentView.do?menuCd=SCD0200273>, 검색일: 2024. 1. 26.

12) 한국특허기술진흥원, "특허분류 연계정보", 한국특허기술진흥원, <https://cls.kipro.or.kr/classification/linkedTable/search>, 검색일: 2024. 1. 26.

13) 정연경, "학문분류표의 재설정에 관한 연구", 「정보관리학회지」, 제17권 제2호(2000), 37-66면.

나누어져야 하며 비특허문헌 조사의 경우에도 한 번에 진행하는 데에 어려움이 있다.

이를 극복하기 위해서, 비특허문헌에도 특허문헌 분류에 사용되는 분류 방법을 적용하여, 특허문헌을 검색하는 것과 같은 방법으로 비특허문헌을 검색할 수 있는 시스템을 제안하는 것이 목표이다. 비특허문헌과 특허문헌이 같은 분류 체계를 가진다면, 비특허문헌 선행기술조사를 위해 들이는 시간이 줄어들 것이고, 반대로 분류된 비특허문헌에 대한 정보를 통해서 특정 비특허문헌에 관련된 특허 문헌을 검색할 때도 사용할 수 있다.

하지만 비특허문헌에 직접 특허분류를 적용하는 것은 시간적으로나 비용적으로 매우 어려운 일이다. 따라서 최근 선행기술조사에서 매우 중요한 것으로 보이며 주요 특허청들이 많은 예산과 인력을 투입하고 있는, 특허문헌에 특허분류 코드를 추천하거나 자동으로 분류해 주는 기계학습 방법[14]을 비특허문헌에 적용하는 것이 효과적인지를 검토하고자 한다. 본 연구에서는 이를 위해서, 기계학습 알고리즘을 이용하여 비특허문헌인 논문에 특허문헌에 사용되는 IPC 코드를 자동분류하는 과정을 예시로 든다. 논문에 IPC 코드가 부여된 데이터베이스가 있다면, IPC 코드 하나로 특허문헌뿐만 아니라 논문 또한 검색할 수 있을 것이며, 논문 외의 비특허문헌들에 IPC 코드 외의 특허분류가 적용된다면 선행기술조사가 효율적으로 이루어질 것으로 예측한다.

1.2. 텍스트 마이닝에 관한 선행 모델

1.2.1. TF-IDF 벡터화

'TF-IDF'는 'Term Frequency - Inverse Document Frequency'로 여러 문서로 이루어진 문서 군이 있을 때 어떤 단어가 특정 문서 내에서 얼마나 중요한 것인지를 나타내는 통계적 수치로, 문서의 핵심어를 추출하는 용도로 사용할 수 있다. 'TF'는 'Term Frequency'로 단어 빈도를 뜻하며, 특정한 단어가 문서 내에서 얼마나 자주 등장하는 지를 나타내는 값이다. 예를 들어, $tf(d,t)$는 특정 문서 d에서의 특정 단어 t의 등장 횟수를 나타낸다.

'IDF'는 'Inverse Document Frequency'로 역문서빈도를 뜻하며, DF(문서빈도, Document Frequency)인 특정한 단어가 문서군 내에서 얼마나 자주 등장하는 지를 나타내는 값의 역수가 'IDF'이며, 아래와 같이 계산한다.

$$idf(t, D) = \log\left(\frac{D}{1 + df(t)}\right)^{15)} \quad (1)$$

여기에서 D는 총 문서의 수를, $df(t)$는 특정 단어 t가 등장한 문서의 수를 나타낸다.

이렇게 계산한 'TF'와 'IDF'를 곱한 값이 'TF-IDF'가 된다.

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D) \quad (2)$$

결과적으로 특정 문서 내에서 단어 빈도가 높을수록, 그리고 전체 문서 중 그 단어를 포함한 문서가 적을수록 'TF-IDF' 값이 커진다.

TF-IDF 벡터화는 이렇게 계산해 낸 'TF-IDF'라는 특정한 값을 사용해서 데이터의 특징을 추출하는 방법이다.[16] 기계학습 모델은 텍스트를 바로 사용할 수 없으므로 문자를 의미가 있는 숫자 값인 벡터로 변환해서 사용한다.[17] 'TF-IDF'는 특정 문서에서 단어의 중요성을 나타내는

14) 심우철 외 4인, "한국 특허문헌 특성 및 딥러닝 기반 분류모델을 고려한 CPC 자동분류에 관한 연구", 한국정보과학회 학술발표논문집, 한국정보과학회, 2020, 406면.

15) 단순히 역수를 취한다면, 총 문서의 수(n)이 커질수록, IDF 값이 기하급수적으로 커지기 때문에, 전체 문서의 수를 해당 단어를 포함한 문서의 수로 나눈 뒤 로그를 취한다.

16) 전창욱 외 3인, 「텐서플로 2와 머신러닝으로 시작하는 자연어 처리」, 위키북스, 2022, 60면.

데 사용되며, 이 값을 사용하면 단순 횟수를 이용하는 것보다 각 단어의 특성을 좀 더 잘 반영할 수 있다.[18]

1.2.2. SBERT

'BERT(Bidirectional Encoder Representations from Transformers)'란 구글이 개발한 자연어처리 사전 훈련 기술로 2019년 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[19]'이라는 논문으로 공개되었으며, 좋은 성능으로 자연어 처리 세계에 혁명을 일으켰었다.[20] 'BERT'에 적용된 아키텍처는 'Transformer'이며 2017년 구글이 발표한 'Attention Is All You Need[21]'라는 논문으로 공개되었다. 'Transformer'는 'BERT' 외에도, 'OpenAI'의 'ChatGPT'에 사용된 'GPT'에서도 사용되며 자연어 처리에서 최고의 기법으로 자리 잡고 있다.

'Transformer'의 원리로는, 먼저 인코더-디코더 프레임워크가 있다. 인코더는 입력 문장의 표현 방법을 학습시키고 그 결과를 디코더로 보낸다. 디코더는 인코더에서 학습한 표현 결과를 입력받아 사용자가 원하는 문장을 생성한다.[22] 이 과정에서 셀프 어텐션이라는 특수한 형태의 어텐션이 적용된다. 셀프 어텐션은 입력 텍스트 내의 어떤 단어가 다른 단어들과 가지는 연결 관계를 파악하여, 단어가 문장 내에서 갖는 의미를 이해하여 문맥 정보를 익힌다.[23] 그리고, 하나의 문제를 해결하고 이와 다르면서 관련된 문제에 적용하는 동안 얻은 지식을 저장하는 데에 집중하는 전이 학습이 적용되었다. 'BERT'는 'Masked Language Model[24]'을 영어 위키피디아에서 사전에 학습하여 적용하였다.

'BERT'는 문맥을 고려한 임베딩 모델이기 때문에 자연어 처리 분야에 크게 이바지해 왔는데[25], 원리는 다음과 같다. 먼저, 'Transformer'의 인코더-디코더 프레임워크와 달리 입력 문장의 표현 방법을 학습하는 인코더만 사용한다. 그리고, 멀티 헤드 어텐션을 사용하여 문장에서 각 단어의 문맥을 이해한 후, 문장에 있는 각 단어의 문맥 표현을 출력으로 반환한다.[26] 멀티 헤드 어텐션은 셀프 어텐션을 여러 번 적용한 것으로, 이를 통해 문맥 정보를 더 잘 파악할 수 있다.

SBERT는 'BERT'의 문장 임베딩 성능을 우수하게 개선한 모델[27]로, 2019년 'Sentence-bert: Sentence embeddings using siamese bert-networks[28]'라는 논문으로 공개되었다. 'BERT'의 문장 임베딩을 응용하여 'BERT'를 파인 튜닝한 것으로, 문장 쌍 분류 태스크(Natural Language Inferencing, NLI) 문제와 문장 쌍 회귀 태스크(Semantic Textual

17) 안상준 외 1인, 「딥 러닝을 이용한 자연어 처리 입문 2권」, 위키독스, 2023, 66면.
18) 전창욱 외 3인, 위의 단행본, 62면.
19) Jacob Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *arxiv*, (2018), pp. 1-16.
20) 수다르산 라비찬디란, 「구글 BERT의 정석」, 한빛미디어, 2022, 7면.
21) Ashish Vaswani et al., "Attention Is All You Need", *arxiv*, (2017), pp. 1-15.
22) 수다르산 라비찬디란, 위의 단행본, 22면.
23) 수다르산 라비찬디란, 위의 단행본, 25면.
24) Masked Language Model(MLM)은 어떤 문장이 있을 때 문장의 특정 부분을 Masking 처리하여 모델이 Masking 처리된 부분을 예측하도록 학습시키는 방식이다. 김주동 외 2인, "스마트한 텍스트 분석을 향한 첫걸음, KoreALBERT", SAMSUNG SDS, <https://www.samsungsds.com/kr/insights/techtoolkit_2021_korealbert.html>, 검색일: 2024. 1. 26.
25) 수다르산 라비찬디란, 위의 단행본, 68면.
26) 수다르산 라비찬디란, 위의 단행본, 70면.
27) 안상준 외 1인, 「딥 러닝을 이용한 자연어 처리 입문 3권」, 위키독스, 2023, 202-205면.
28) Nils Reimers & Iryna Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks", *arxiv*, (2019), pp. 1-11.

Similarity, STS) 문제를 푸는 것으로 학습된다.

1.2.3. 코사인 유사도

코사인 유사도(Cosine Similarity)는 내적공간의 두 벡터 간 각도의 코사인값을 이용하여 측정된 벡터 간의 유사한 정도를 의미한다. 이는 텍스트 마이닝 분야에서, 단어 하나하나는 각각의 차원을 구성하고 문서는 각 단어가 문서에 나타나는 횟수로 표현되는 벡터값을 가질 때 두 문서의 유사도를 측정하는 매우 유용한 방법이다.[29] 코사인 유사도를 계산하는 방법은 다음과 같다.

$$similarity = \cos(\theta) = \frac{A \cdot B}{\| A \| \| B \|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (3)$$

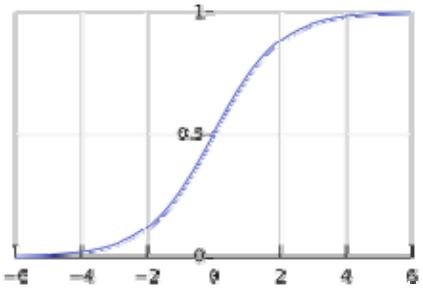
이렇게 계산된 유사도는 −1에서 1까지의 값을 가지며, −1은 서로 완전히 반대되는 경우, 0은 서로 독립적인 경우, 1은 서로 완전히 같은 경우를 의미한다.

1.2.4. Logistic Regression

로지스틱 회귀(Logistic Regression)는 영국의 통계학자인 'D. R. Cox'가 1958년에 제안한 확률 모델[30]로서 독립 변수의 선형 결합을 이용하여 사건의 발생 가능성을 예측하는 데 사용되는 통계 기법이다. 로지스틱 회귀의 목적은 일반적인 회귀 분석의 목표와 같이 종속 변수와 독립 변수 간의 관계를 구체적인 함수로 나타내어 향후 예측 모델에 사용하는 것이다. 이는 독립변수의 선형 결합으로 종속 변수를 설명한다는 관점에서는 선형 회귀 분석과 유사하다. 하지만 로지스틱 회귀는 각 특성에 대해 모델 정확도를 최대화하는 적절한 가중치 또는 계수를 찾는다. 선형 회귀처럼 각 항의 합을 그대로 출력하는 대신 로지스틱 회귀는 시그모이드 함수(Sigmoid Function)를 적용한다.[31] 시그모이드 함수는 아래와 같이 정의되며, 그림 1은 시그모이드 함수의 그래프를 나타낸다.

$$sig(t) = \frac{1}{1 + e^{-t}} \quad (4)$$

<그림 1 시그모이드 함수의 그래프>



29) Singhal Amit, "Modern Information Retrieval: A Brief Overview", *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, Vol.24 No. 4(2001), pp. 35-43.
30) David R. Cox, "The regression analysis of binary sequences (with discussion)", *J Roy Stat Soc B*, Vol.20(1958), pp. 215-242.
31) 코리 웨이드, 「XGBoost와 사이킷런을 활용한 그레이디언트 부스팅」, 한빛미디어, 2022, 64면.

1.2.5. XGBClassifier[32)]

결정 트리(Decision Tree)는 의사 결정 규칙과 그 결과들을 트리 구조로 도식화한 의사 결정 지원 도구의 일종이다. 하지만, 과대 적합 되기 쉬워서[33)] 새로운 데이터에 맞지 않는 기계학습 모델을 만들기 때문에, 배깅과 부스팅과 같은 앙상블 방법을 통해 결정 트리를 연결하는 것이 효과적이다. 부스팅이란 이전 트리의 오차를 기반으로 새로운 트리를 훈련하는 것을 기본적인 아이디어로 둔 알고리즘으로, 가장 효과적이라고 알려진 것이 그레이디언트 부스팅이다. 그레이디언트 부스팅은 잘못된 예측을 기반으로 조정되지만, 이전 트리의 예측 오차를 기반으로 완전히 새로운 트리를 훈련한다.[34)]

워싱턴 대학교의 티엔치 첸(Tianqi Chen)과 카를로스 게스트린(Carlos Guestrin)은 2016년 'XGBoost: A Scalable Tree Boosting System[35)]'이라는 논문에서 그레이디언트 부스팅의 일관성과 성능을 더 향상하며 여기에 익스트림 그레이디언트 부스팅(XGBoost)이라는 이름을 붙였다. 'XGBoost'는 결정 트리를 기초로 하고 있으므로, 회귀 모델과 분류 모델을 가지고 있다. 이중, 분류모델을 사용하기 위한 클래스를 XGBClassifier라고 한다.[36)]

2. 연구 설계

2.1. 선행연구 검토

2.1.1. 특허문헌 자동분류에 관한 선행연구

특허문헌의 특허분류 코드를 추천하거나 자동으로 분류를 해주는 시스템은 선행기술조사에서 매우 중요한 것으로 보이며, 주요 특허청에서는 많은 예산과 인력을 투입하고 있다.[37)]

박찬정 외 3인은 '용어 클러스터링을 이용한 특허문서 자동 IPC 분류[38)]'에서 기계학습을 이용한 IPC 자동분류에 관한 연구를 수행했다. 특허문헌에 사용된 명사들 사이에 친밀도라는 개념을 소개하고, 이를 이용한 용어 클러스터링을 사용하여 자동분류에 드는 학습 시간을 단축하고 정확도를 높이는 방안을 제시하였다.

임소라/권용진은 '특허문서 필드의 기능적 특성을 활용한 IPC 다중 레이블 분류[39)]'에서 특허분류 과정을 고찰한 기계학습 IPC 자동분류에 관한 연구를 수행했다. 특허문헌을 분류할 때 영향을 미치는 기술 분야 및 배경 기술 필드를 활용한 다중 레이블 분류 모델을 구축하여, 특허문헌의 데이터 구조를 바탕으로 하여 정확도를 높이는 방안을 제시하였다.

심우철 외 4인은 '한국 특허문헌 특성 및 딥러닝 기반 분류모델을 고려한 CPC 자동분류에 관한 연구[40)]'에서 한국 특허문헌의 특성을 반영한 기계학습 CPC 자동분류에 관한 연구를 수행했다. 최근 자연어처리 분야에서 뛰어난 성능을 보이는 pre-trained BERT 모델을 이용하였고,

32) Jiaming Yuan, "xgboost", GitHub, <https://github.com/dmlc/xgboost>, 검색일: 2024. 1. 26.
33) 코리 웨이드, 위의 단행본, 70면.
34) 코리 웨이드, 위의 단행본, 126-127면.
35) Tianqi Chen & Carlos Guestrin, "Xgboost: A scalable tree boosting system", *arxiv*, (2016), pp. 1-13.
36) Jiaming Yuan, "XGBoost", DMLC XGBOOST, <https://xgboost.readthedocs.io/en/stable/index.html>, 검색일: 2024.01.26.
37) 심우철 외 4인, 위의 학술대회 논문집, 406면.
38) 박찬정 외 3인, "용어 클러스터링을 이용한 특허문서 자동 IPC 분류", 「한국정보기술학회논문지」, 제12권 제9호(2014), 127-139면.
39) 임소라/권용진, "특허문서 필드의 기능적 특성을 활용한 IPC 다중 레이블 분류", 「인터넷정보학회논문지」, 제18권 제1호(2017), 77-88면.
40) 심우철 외 4인, 위의 학술지, 406-408면.

특허문헌 필드의 다양한 조합에 적용하여 정확도를 높이는 방안을 제시하였다.

박진우 외 4인은 '한국어 특허 문장 기반 CPC 자동분류 연구 ―인공지능 언어모델 KorPatBERT를 활용한 딥러닝 기법 접근―[41]'에서 인공지능 언어모델인 BERT 모델을 이용한 CPC 자동분류에 관한 연구를 수행했다. 기존 모델 대비 우수한 KorPatBERT 모델을 제시하고, CPC 코드의 불균형적인 데이터 분포를 완화한 학습데이터 세트를 제안하여 정확도를 높이는 방안을 제시하였다.

2.1.2. 비특허문헌 자동분류에 관한 선행연구

김현종 외 3인은 '비지도학습 기반의 행정부서별 신문기사 자동분류 연구[42]'에서 신문 기사 자동분류에 관한 연구를 수행했다. Word2Vec와 토픽모델링 기법을 적용하는 방식과, 비지도 학습 기반의 자동분류 방식을 제시하였다.

오석범/강현국은 '머신러닝 알고리즘을 이용한 표준문서의 ICS 코드 자동분류(ICAC)[43]'에서 표준문서의 ICS 코드 자동분류에 관한 연구를 수행했다. GRU 알고리즘을 비롯한 여러 알고리즘을 비교·분석하여 정확도를 높이는 방안을 제시하였다.

김판준은 '자질선정을 통한 국내 학술지 논문의 자동분류에 관한 연구[44]'에서 국내 학술지 논문에 학술연구분야분류 표상의 분류 범주를 자동 할당하는 연구를 수행했다. 자동분류에 자질 순위화 기법을 적용하여 정확도를 높이는 방안을 제시하였다.

하지만 기존 문헌들에서 특허분류 체계를 비특허문헌에 적용한 연구는 없었다. 본 연구에서는 특허문헌 텍스트 마이닝을 통한 비특허문헌 자동분류와 관련하여, 특허문헌 텍스트 마이닝을 통해서 특허문헌 자동분류에 사용되는 기계학습 알고리즘을 비특허문헌에 적용한다. 이를 통해서 비특허문헌에 특허분류 체계를 적용한 예측 결과를 확인하여 연구 목표의 효율성을 검증한다.

2.2. 연구 방법

2.2.1. 개요

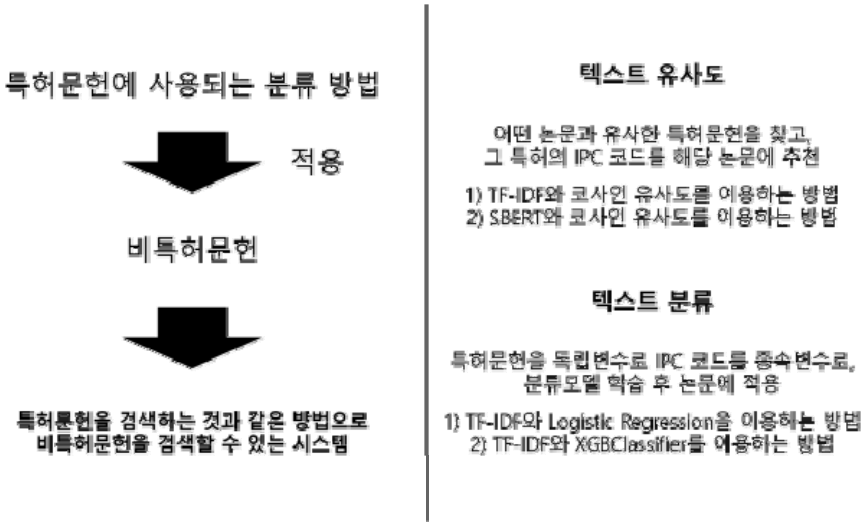
특허문헌 텍스트 마이닝을 통해서 비특허문헌을 자동분류하기 위해서 2가지 프로세스를 제안한다. 첫 번째는 텍스트 유사도 측정 알고리즘을 사용하는 방법으로, 먼저 어떤 비특허문헌과 유사한 여러 가지 특허문헌을 찾는다. 이후, 비특허문헌과 유사하다고 판단된 특허들을 살펴보고 해당 특허들의 특허분류 코드를 비특허문헌에 추천하는 방식으로 자동분류를 구현할 수 있다. 두 번째는 텍스트 분류 알고리즘을 사용하는 방법으로, 먼저 여러 가지 특허문헌을 독립변수로 하고 각 특허문헌의 메인 특허분류 코드를 종속변수로 하여 기계학습을 이용한 모델을 만든다. 이후, 해당 모델을 비특허문헌에 적용하면 많은 비특허문헌에 특허분류 코드를 추천할 수 있고 자동분류를 구현할 수 있다. 전체적인 연구 절차는 그림 2와 같다.

본 연구에서는 텍스트 유사도 측정 알고리즘을 사용하는 방법과 텍스트 분류 알고리즘을 사

41) 박진우 외 4인, "한국어 특허 문장 기반 CPC 자동분류 연구 ―인공지능 언어모델 KorPatBERT를 활용한 딥러닝 기법 접근―", 「지식재산연구」, 제17권 제3호(2022), 209-256면.
42) 김현종 외 3인, "비지도학습 기반의 행정부서별 신문기사 자동분류 연구", 「한국산학기술학회논문지」, 제21권 제9호(2020), 345-351면.
43) 오석범/강현국, "머신러닝 알고리즘을 이용한 표준문서의 ICS 코드 자동분류(ICAC)", 「표준인증안전학회지」, 제11권 제2호(2021), 157-173면.
44) 김판준, "자질선정을 통한 국내 학술지 논문의 자동분류에 관한 연구", 「정보관리학회지」, 제39권 제1호(2022), 69-90면.

용하는 방법을 제안하기 위해서, 특허문헌의 요약 부분과 비특허문헌인 논문의 요약 부분을 예시로 든다. 또한, 텍스트 분류 알고리즘을 사용하는 방법을 적용할 때, IPC 코드 전체를 사용한다면 각 분류별 데이터의 양이 절대적으로 부족해지므로 전체를 사용하지 않고 섹션, 클래스, 서브 클래스까지만을 사용했다.[45] 그림3을 살펴보았을 때, 전체 사용된 40828개의 IPC코드 중에서 섹션만을 분류하였을 때 평균 약 5100개의 코드가 분류되었고, 클래스까지 분류하였을 때 평균 약 523개, 서브클래스까지 분류하였을 때 평균 약 178개, 메인그룹까지 분류하였을 때 평균 약 52개, 서브그룹까지 모두 분류하였을 때 평균 약 13개로 분류되었다.

<그림 2 연구 절차>



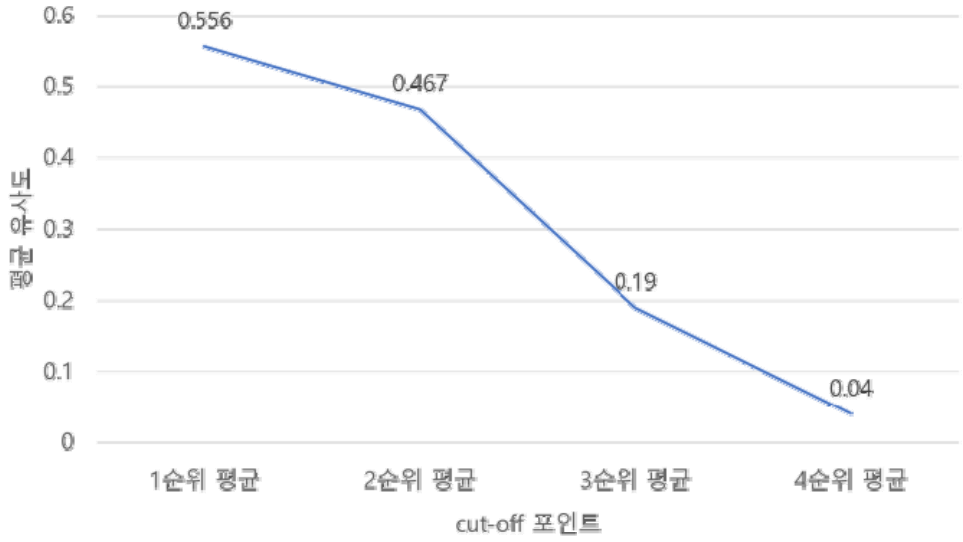
<그림 3 각 분류별 데이터의 양>

45) 박진우 외 4인, 위의 학술지, 237-240면.

2.2.2. 텍스트 유사도 측정 알고리즘을 사용하는 방법

코사인 유사도는 많은 차원 공간에서도 거리를 측정할 수 있으므로 문서의 유사도를 비교하기 위해서 널리 사용된다. 이를 계산하기 위해서는, 단어나 문장을 벡터로 변환하는 과정이 필요하므로, 본 연구에서는 이를 위해서 TF-IDF를 이용하는 방법과 SBERT를 이용하는 방법을 예시로 든다. TF-IDF는 각 단어의 특성을 잘 반영할 수 있으므로 흔히 사용되고, SBERT는 BERT로부터 문장 임베딩을 얻을 수 있는 Sentence Transformer를 말하며[46] 임베딩 분야에서 최근 우수한 성능을 보인다. 두 모델을 통해 특허문헌과 비특허 문헌을 벡터화하여 표현하고, 이후, 어떤 비특허분헌과 전체 특허문헌들 사이의 코사인 유사도를 모두 계산하고 순위를 매긴다. 코사인 유사도의 순위를 cut-off 포인트로 사용할 경우, 순위는 높지만 유사도가 낮은 경우가 생기는 단점이 있다. 하지만 해당 실험에서는 적절한 임계치를 알 수 없는 상황이므로, 순위를 cut-off 포인트로 사용하는 방법으로 특정 비특허문헌에 적절한 특허분류를 찾아나간다. 또한, 그림 4를 살펴보았을 때, 3순위에서의 평균 유사도부터 급격하게 하락하므로 2순위까지의 문헌을 살펴보았다. 이를 통해 비특허문헌과 유사한 특허문헌을 찾아 해당 특허문헌의 IPC 코드를 확인하고 추천할 수 있다.

<그림 4 Cut-off 포인트 설정>



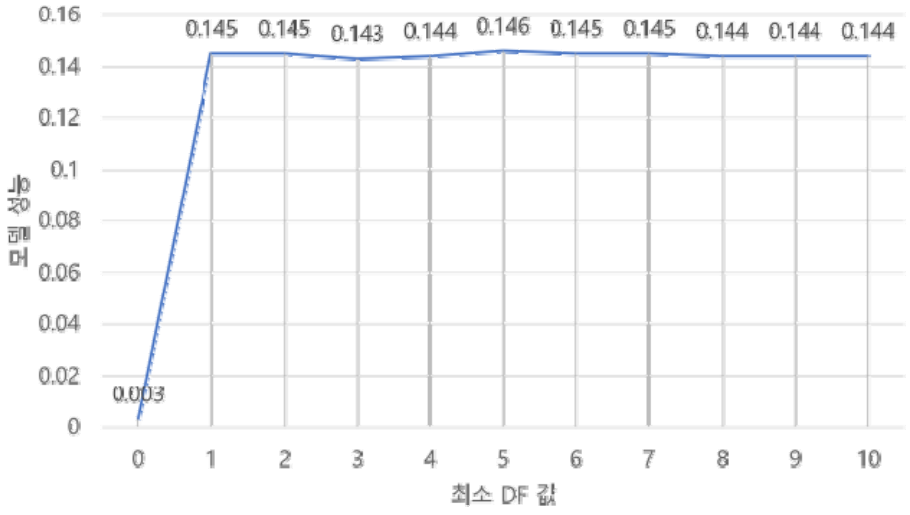
2.2.3. 텍스트 분류 알고리즘을 사용하는 방법

본 연구에서는 텍스트 분류 알고리즘을 사용하는 방법을 제안하기 위한 2가지 방법을 예시로 든다. 첫 번째는 TF-IDF와 Logistic Regression을 이용하는 방법이다. Logistic Regression은 모델의 개념과 결과 해석이 비교적 쉽고, 이진 분류와 다중 클래스 분류 문제에 모두 적용할 수 있다는 점에서 분류 문제에서 자주 사용된다. TF-IDF 벡터화를 통해서 특허문헌과 비특허문헌을 벡터의 형태로 표현한다. 이후, 특허문헌들의 벡터값들을 독립변수로 하고 각 특허문헌의 메인 특허분류 코드를 종속변수로 하여 Logistic Regression 모델을 학습한다. 마지막으로, 학습한 모델을 비특허문헌들의 벡터값들에 적용하여 각 비특허문헌에 특허분류 코드를 추천한다. 두 번째는 TF-IDF와 XGBClassifier를 이용하는 방법이다. XGBClassifier는 최근 분류 문

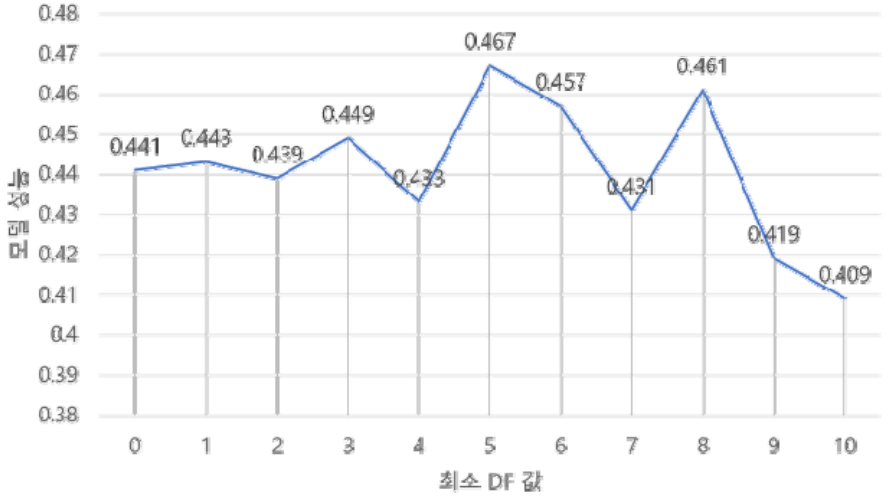
46) Reimers N & Gurevych I, 위의 학술지, pp.1-11.

제에서 우수한 성능을 보이는 모델이다. TF-IDF 벡터화 과정은 TF-IDF와 Logistic Regression을 이용하는 방법과 동일하며, Logistic Regression 모델을 학습하는 대신에 XGBClassifier 모델을 학습한다. 이후, 학습한 모델을 비특허문헌들의 벡터값들에 적용하여 각 비특허문헌에 특허분류 코드를 추천한다. 이를 위해 적절한 하이퍼파라미터를 정하려고, Logistic Regression 모델과 XGBClassifier 모델에서 TF-IDF 성능을 실험하였다. 그 결과 그림 5와 그림 6에서 나타나듯이, 최소 DF의 값을 5로 설정했을 때가 성능이 가장 우수하여 결과를 연구에 반영하였다.

<그림 5 Logistic Regression 모델의 최소 DF 값 설정 실험>



<그림 6 XGBClassifier 모델의 최소 DF 값 설정 실험>

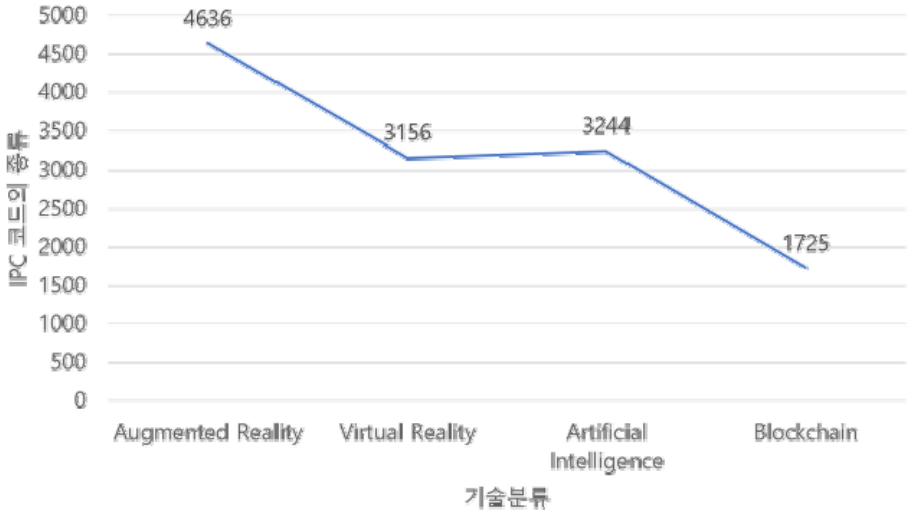


2.3. 연구 범위

2.3.1. 특허문헌 연구 범위

특허문헌 텍스트 마이닝을 위하여, 윕스(WIPS)에서 제공하는 윈텔립스(WINTELIPS)를 이용하였다. 원활한 실험을 위해 문헌의 양을 줄이고자, 2023년 8월 19일 기준으로 미국의 등록특허 중 명칭, 요약, 대표청구항에 “augmented reality”와 정확히 일치하는 구문이 있는 문헌을 검색하였다. 그림 7을 살펴보았을 때, 제4차 산업 혁명을과 관련된 기술 중, AR 기술이 가장 많은 IPC 코드의 종류를 가지고 있었다. 다른 키워드로도 적절한 양의 데이터를 얻을 수 있지만, 많은 IPC 코드의 종류를 가지고 있는 기술을 연구의 범위로 설정했을 때, 텍스트 유사도 측정 알고리즘과 텍스트 분류 알고리즘에서 다양한 데이터를 학습하여 효과적인 결과를 도출해낼 수 있을 것으로 판단했다. 검색된 특허 수는 6984건으로 ‘발명의 명칭’, ‘요약’, ‘대표청구항’, ‘Current IPC All’의 서지사항을 csv 형태로 저장하였다.

<그림 7 검색 키워드 설정>



<표 1 특허문헌 연구 범위>

자료구분	국가	검색DB	검색구간	검색범위	컴색키워드	특허수
등록특허	미국	WINTELIPS	전체	명칭, 요약, 대표청구항	“augmented reality”	6984

2.3.2. 비특허문헌 연구 범위

비특허문헌 자동분류를 위하여, 엘스비어(ELSEVIER)에서 제공하는 스코퍼스(SCOPUS)를 이용하였다. 원활한 실험을 위해 문헌의 양을 줄이고자, 2023년 8월 19일 기준으로 2022년의 논문 중 제목, 요약문, 키워드에 “augmented reality”와 정확히 일치하는 구문이 있는 문헌을 검색하였다. 검색된 논문 수는 5911건으로 ‘Title’, ‘Abstract’, ‘Index Keywords’의 서지사항을 csv 형태로 저장하였다.

<표 2 비특허문헌 연구 범위>

자료구분	국가	검색DB	검색구간	검색범위	컴색키워드	논문수
논문	전체	SCOPUS	2022년	제목, 요약문, 키워드	"augmented reality"	5911

3. 비특허문헌의 자동분류 방법

3.1. 텍스트 유사도 측정 알고리즘을 사용하는 방법

3.1.1. TF-IDF와 코사인 유사도를 이용하는 방법

특허문헌과 비특허문헌을 TF-IDF로 벡터화하기 위해서, 아나콘다를 설치한 후 가상환경을 만들어 파이썬 버전 3.6.13에서 Scikit-learn 라이브러리[47] 문서전처리 기능의 TfidfVectorizer 클래스를 이용한다. 다음 단계로, IPC 코드를 추천받고 싶은 어떤 비특허문헌을 선택한다. 선택한 비특허문헌과 모든 특허문헌(6984개) 간의 코사인 유사도를 계산하고 내림차순으로 정렬한다. 이후, 순위권 상위에 있는 특허문헌들의 내용과 IPC 코드를 확인하고 추천한다. 번역은 구글 번역기를 사용했으며, 결과는 다음과 같다.

예시 논문 1번의 요약의 번역

> 에너지에 대한 대외 의존 문제와 국내/국제 환경 정책 방향은 재생 에너지 시스템에 대한 투자를 장려합니다. 신재생에너지시스템(RES)에서는 태양에너지를 기반으로 한 태양광발전(PV) 시스템이 매년 증가하고 있다. 이러한 상황은 작업 영역이 태양광 발전 시스템인 자격을 갖춘 직원의 필요성을 보여줍니다. 본 연구에서는 PV 전력 시스템의 산업 보건 및 안전(OHS) 문제에 대한 문헌 검색을 수행했습니다. 동시에 현장 연구와 전문가 인터뷰를 실시하여 업무 영역에서 받은 OHS 관행과 교육을 조사했습니다. 이러한 방향에서 PV 전력 시스템의 OHS 성공을 위한 OHS 요구 사항 및 교육의 중요성이 강조되었습니다. PV 전력 시스템 분야에서 OHS의 성과와 성공을 높이려면 교육의 중요성이 강조되었습니다. 본 연구에서는 PV를 설치하는 동안 고소 작업을 하는 두 가지 다른 활동에 대한 샘플 게임화 시나리오를 만들었습니다. 생성된 샘플 게임화 시나리오를 사용하여 OHS 교육에서 VR/AR 애플리케이션을 개선하고 이러한 교육 애플리케이션의 보급을 지원하기 위한 알고리즘이 개발되었습니다. 이렇게 VR/AR 기술에 필요한 정보를 학문적 차원에서 설명하고, 지속가능성 관점에서 콘텐츠 개발자, 연구자, 관련 기관 및 단체에 기여하는 것을 목표로 합니다.

예시 논문 1번과의 유사도가 1위인 특허 요약의 번역과 IPC 코드

> 태양광(PV) 태양광 현장 특정 활동 환경과 관련된 정보가 별도의 컴퓨팅 장치나 장치가 아닌 주변 환경의 일부로 증강 현실(AR) 장치의 디스플레이 화면에 표시되는 시스템 및 방법 설치 매뉴얼. 활동 환경에는 태양광발전소 부지별 조사 및 타당성 분석, 설치 및 시운전, 운영 및 유지관리(O&M), 부지 점검/철거 활동이 포함됩니다. 따라서 설치자/기술자는 관련 정보를 손쉽게 사용할 수 있으므로 설치자/기술자는 집중력을 잃지 않고 관련된 작업을 계속할 수 있습니다. AR 장치는 현장 조사 중에 관련 데이터를 기록하고 중요한 사항을 기록하며, 설치 및 시운전을 가속화하고, O&M을 보다 효율적으로 만들고, 현장의 지속적인 개선/관리를 위한 전체 프로세스를 기록할 수 있습니다. 또한 전체 설치 프로그램에 걸쳐 프로세스를 균일하게 만들기 위한 지침으로 회사별 모범 사례를 AR 장치에 로드할 수 있습니다.
>
> IPC 코드 : G06T-019/00, G09G-005/00, G06K-009/00, G02B-027/01, G06T-011/60

47) 파이썬 프로그래밍 언어용 자유 소프트웨어 기계 학습 라이브러리이다.

예시 논문 1번과의 유사도가 2위인 특허 요약의 번역과 IPC 코드

홀로그램, 라이트 필드, 가상, 증강 및 혼합 현실 애플리케이션을 위한 홀로그램 불투명 변조 상태의 중첩을 위한 투명 에너지 릴레이 도파관 시스템이 개시됩니다. 광 필드 시스템은 하나 이상의 에너지 변조 요소를 갖는 하나 이상의 에너지 도파관 중계 시스템을 포함할 수 있으며, 각 에너지 변조 요소는 이를 통과하는 에너지를 변조하도록 구성되어, 이를 통해 통과하는 에너지는 4D 플렌옵틱 함수 또는 그 역에 따라 지향될 수 있습니다.

IPC 코드 : H04N-023/957, G02B-030/00, G02B-030/33, G02B-030/56, H04N-013/388, H04N-013/344, G06F-003/01, F21V-008/00, G02B-006/02, G02B-006/04, G02B-006/08, G02B-006/293, G02B-027/00, G02B-027/01, G02B-027/09, G02B-027/10, G02B-003/00, G02B-003/08, G02B-005/32, G02B-025/00, G03H-001/00, G03H-001/02, G03H-001/22, G10K-011/26, G21K-001/00, H04N-005/89

예시 논문 2번의 요약의 번역

유산에서 스마트 글래스 증강 현실(AR)의 적용이 증가하고 있음에도 불구하고 의미 있고 교육적인 몰입형 유산 경험을 설계하기 위한 기반이 될 수 있는 프레임워크는 없습니다. 이 기사에서는 정서적 경험을 학습과 연결하고 비교훈적인 스토리텔링 매체인 AR을 실제로 탐색하는 문헌을 활용하여 유적지에서의 AR 경험을 위한 프로토타입 디자인 프레임워크를 제안합니다. 여기에서 스마트 글래스 AR은 정서적 상호 작용을 생성하기 위한 중요한 기술 이정표로 간주됩니다. 이는 방문자/시청자에게 현지화된 과거를 경험하고, 구현하고, 물리적, 사회적 상호 작용을 하고 이에 대해 배울 수 있는 새로운 방법을 제공하는 것입니다.

예시 논문 2번과의 유사도가 1위인 특허 요약의 번역과 IPC 코드

정보를 3D 요소로 표시하는 스마트 글래스, 네비게이션 보조 부품 및 중앙 데이터베이스를 포함하는 차량 사용자에게 정보를 표시하는 시스템이 제공되며, 상기 스마트 글래스와 네비게이션 보조 장치는 구성요소는 중앙 데이터베이스와 통신 중입니다. 이 데이터베이스는 정보를 저장하고 차량이 조종되는 위치에 따라 관련 업데이트 정보를 스마트 글래스에 실시간으로 전송하고 표시하는 데 도움을 줍니다. 또한, 마커리스 증강현실(AR) 기술을 통해 정보가 스마트 글래스에 표시되므로 정보를 표시하기 위한 외부 장치나 프로젝터가 필요하지 않습니다.

IPC 코드 : G06F-003/14, G01C-021/36, G02B-027/01, H04N-005/33

예시 논문 2번과의 유사도가 2위인 특허 요약의 번역과 IPC 코드

설명된 증강 현실(AR) 시스템 및 장치는 디지털 콘텐츠를 사용하여 사용자에게 향상된 인간 감각 인식을 제공합니다. 특히, 스마트글래스는 물체 식별 및 광학 흐름 추적을 사용하여 문화 유적지 방문자에게 몰입형 AR 경험을 선사합니다. 시스템 소프트웨어 플랫폼과 방법론은 사용자 웨어러블 기기를 통해 문화 현장 방문자에게 몰입형 AR 경험을 설계하고 배포하는 데 특히 적합합니다. 서투른 휴대용 기기의 방해 없이 사용자는 증강 환경, 실내 및 실외를 장벽 없이 자유롭게 돌아다니며 이야기를 눈앞에서 펼쳐볼 수 있습니다. 문화 유적지가 살아나면서 유적지에 대한 교육 성과가 향상됩니다. 그래픽 사용자 인터페이스를 사용하는 이 시스템을 통해 플랫폼 관리자는 서로 다른 문화 사이트 콘텐츠를 팔림프세스트(palimpsest) 또는 투어라고 하는 일관된 AR 스토리텔링 경험으로 변환할 수 있습니다. 학습 단계에서 관심 지점이 목록화되고 다양한 각도와 다양한 조명 조건에서 이미지가 촬영되어 마커가 생성됩니다. 위치정보 시스템을 사용하면 데이터가 한 번에 한 공간씩 스마트글래스에 로드되어 효율성이 향상됩니다. 사용 시 유사성 및 임계값 알고리즘을 사용하여 관심 지점 마커를 일치시킨 후 AR 콘텐츠가 생성됩니다. 광학 흐름 추적은 사용자 움직임을 추적하고 AR 경험을 향상시키는 데 사용됩니다.

IPC 코드 : G06V-020/20, G06T-007/246, G06T-007/73, G06Q-010/00, G06Q-010/06, G02B-027/01, G06F-003/01

3.1.2. SBERT와 코사인 유사도를 이용하는 방법

특허문헌과 비특허문헌을 SBERT로 임베딩하여 벡터의 형태로 표현하기 위해서, 아나콘다를 설치한 후 가상환경을 만들어 파이썬 버전 3.10.9에서 sentence_transformers 라이브러리를 이용한다. 다음 단계로, IPC 코드를 추천받고 싶은 어떤 비특허문헌을 선택한다. 선택한 비특허문헌과 모든 특허문헌(6984개) 간의 코사인 유사도를 계산하고 내림차순으로 정렬한다. 이후, 순위권 상위에 있는 특허문헌들의 내용과 IPC 코드를 확인하고 추천한다. 번역은 구글 번역기를 사용했으며, 결과는 다음과 같다.

예시 논문 1번의 요약의 번역

에너지에 대한 대외 의존 문제와 국내/국제 환경 정책 방향은 재생 에너지 시스템에 대한 투자를 장려합니다. 신재생에너지시스템(RES)에서는 태양에너지를 기반으로 한 태양광발전(PV) 시스템이 매년 증가하고 있다. (이하 생략)

예시 논문 1번과의 유사도가 1위인 특허 요약의 번역과 IPC 코드

태양광(PV) 태양광 현장 특정 활동 환경과 관련된 정보가 별도의 컴퓨팅 장치나 장치가 아닌 주변 환경의 일부로 증강 현실(AR) 장치의 디스플레이 화면에 표시되는 시스템 및 방법 설치 매뉴얼. 활동 환경에는 태양광발전소 부지별 조사 및 타당성 분석, 설치 및 시운전, 운영 및 유지관리(O&M), 부지 점검/철거 활동이 포함됩니다. 따라서 설치자/기술자는 관련 정보를 손쉽게 사용할 수 있으므로 설치자/기술자는 집중력을 잃지 않고 관련된 작업을 계속할 수 있습니다. AR 장치는 현장 조사 중에 관련 데이터를 기록하고 중요한 사항을 기록하며, 설치 및 시운전을 가속화하고, O&M을 보다 효율적으로 만들고, 현장의 지속적인 개선/관리를 위한 전체 프로세스를 기록할 수 있습니다. 또한 전체 설치 프로그램에 걸쳐 프로세스를 균일하게 만들기 위한 지침으로 회사별 모범 사례를 AR 장치에 로드할 수 있습니다. IPC 코드 : G06T-019/00, G09G-005/00, G06K-009/00, G02B-027/01, G06T-011/60

예시 논문 1번과의 유사도가 2위인 특허 요약의 번역과 IPC 코드

실제 또는 시뮬레이션 열화상 장비와 함께 증강 현실(AR)을 사용하는 방법입니다. 주요 응용 프로그램은 응급 상황 대처자를 교육하는 것입니다. 특히 열화상 카메라나 일반 비디오 카메라, 추적 시스템을 활용해 AR 환경에서 추적 시점을 제공한다. 환경의 증강된 부분은 화재, 연기, 소화제 또는 기타 비상 상황으로 구성될 수 있습니다. 이를 통해 열화상을 사용한 소방, 피해 통제, 수색 및 구조 및 기타 최초 대응 기술에 대한 저렴하고 유연하며 현실적인 현장 교육이 가능합니다. IPC 코드 : G09G-005/00

예시 논문 2번의 요약의 번역

유산에서 스마트 글래스 증강 현실(AR)의 적용이 증가하고 있음에도 불구하고 의미 있고 교육적인 몰입형 유산 경험을 설계하기 위한 기반이 될 수 있는 프레임워크는 없습니다. 이 기사에서는 정서적 경험을 학습과 연결하고 비교훈적인 스토리텔링 매체인 AR을 실제로 탐색하는 문헌을 활용하여 유적지에서의 AR 경험을 위한 프로토타입 디자인 프레임워크를 제안합니다. (이하 생략)

예시 논문 2번과의 유사도가 1위인 특허 요약의 번역과 IPC 코드

설명된 증강 현실(AR) 시스템 및 장치는 디지털 콘텐츠를 사용하여 사용자에게 향상된 인간 감각 인식을 제공합니다. 특히, 스마트글래스는 물체 식별 및 광학 흐름 추적을 사용하여 문화 유적지 방문자에게 몰입형 AR 경험을 선사합니다. 시스템 소프트웨어 플랫폼과 방법론은 사용자 웨어러블 기기를 통해 문화 현장 방문자에게 몰입형 AR 경험을 설계하고 배포하는 데 특히 적합합니다. 서투른 휴대용 기기의 방해 없이 사용자는 증강 환경, 실내 및 실외를 장벽 없이 자유롭게 돌아다니며 이야기를 눈앞에서 펼쳐볼 수 있습니다. 문화 유적지가 살아나면서 유적지에 대한 교육 성과가 향상됩니다. 그래픽 사용자 인터페이스를 사용하는 이 시스템을 통해 플랫폼 관리자는 서로 다른 문화 사이트 콘텐츠를 팔림프세스트(palimpsest) 또는 투어라고 하는 일관된 AR 스토리텔링 경험으로 변환할 수 있습니다. 학습 단계에서 관심 지점이 목록화되고 다양한 각도와 다양한 조명 조건에서 이미지가 촬영되어 마커가 생성됩니다. 위치정보 시스템을 사용하면 데이터가 한 번에 한 공간씩 스마트글래스에 로드되어 효율성이 향상됩니다. 사용 시 유사성 및 임계값 알고리즘을 사용하여 관심 지점 마커를 일치시킨 후 AR 콘텐츠가 생성됩니다. 광학 흐름 추적은 사용자 움직임을 추적하고 AR 경험을 향상시키는 데 사용됩니다. IPC 코드 : G06V-020/20, G06T-007/246, G06T-007/73, G06Q-010/00, G06Q-010/06, G02B-027/01, G06F-003/01

예시 논문 2번과의 유사도가 2위인 특허 요약의 번역과 IPC 코드

최근 기술 발전으로 인해 가상 현실(VR) 및 증강 현실(AR) 시스템의 범위, 범위 및 경제성이 확대되었습니다. 이전보다 더 많은 사람들이 VR과 AR 시스템을 사용할 수 있게 되었습니다. 그러나 이러한 시스템을 위한 휴대용 물리적 제어 장치는 아직 비슷한 발전을 이루지 못했습니다. 이러한 이유로 개발자가 인간의 능력을 최대한 활용하는 애플리케이션을 만드는 것은 여전히 어렵습니다. 이 제안에서는 VR/AR 시스템을 제어하기 위해 기성 스마트워치를 저렴한 그립 또는 외장 세트와 결합하는 시스템 및 방법을 설명합니다. 우리의 접근 방식을 사용하면 컨트롤러의 모든 계산과 전력이 스마트워치 장치에서 파생되므로 그립이 거의 모든 형태를 취할 수 있으며 개발자와 디자이너가 훨씬 더 다양한 상호 작용 스타일을 애플리케이션에 통합할 수 있습니다. IPC 코드 : G06T-019/00, G06F-003/01, G02B-027/01, G06F-001/16, G06K-009/00, H04N-005/232

3.2. 텍스트 분류 알고리즘을 사용하는 방법

3.2.1. TF-IDF와 Logistic Regression을 이용하는 방법

특허문헌과 비특허문헌을 TF-IDF로 벡터화하기 위해서, 아나콘다를 설치한 후 가상환경을 만들어 파이썬 버전 3.6.13에서 Scikit-learn 라이브러리 문서전처리 기능의 TfidfVectorizer 클래스를 이용한다. 다음 단계로, 특허문헌들의 벡터값들을 독립변수로 하고 각 특허문헌의 메인 특허분류 코드를 종속변수로 하여 Logistic Regression 모델을 학습하기 위해서, 아나콘다를 설치한 후 가상환경을 만들어 파이썬 버전 3.6.13에서 Scikit-learn 라이브러리의 LogisticRegression 클래스를 이용한다. 해당 클래스를 통해서 다중 Logistic Regression 또한 구현 가능하다. 이후, 학습한 모델을 비특허문헌들의 벡터값들에 적용하여 각 비특허문헌에 특허분류 코드를 추천한다. 결과는 다음과 같다.

예시 논문 1번의 요약의 번역

에너지에 대한 대외 의존 문제와 국내/국제 환경 정책 방향은 재생 에너지 시스템에 대한 투자를 장려합니다. 신재생에너지시스템(RES)에서는 태양에너지를 기반으로 한 태양광발전(PV) 시스템이 매년 증가하고 있다. (이하 생략)

예시 논문 1번에 추천된 IPC 코드

IPC 코드 : G06T G 섹션 : 물리학 G06 : 전산; 계산 또는 계수 G06T : 이미지 데이터 처리 또는 발생, 일반

예시 논문 2번의 요약의 번역

유산에서 스마트 글래스 증강 현실(AR)의 적용이 증가하고 있음에도 불구하고 의미 있고 교육적인 몰입형 유산 경험을 설계하기 위한 기반이 될 수 있는 프레임워크는 없습니다. 이 기사에서는 정서적 경험을 학습과 연결하고 비교훈적인 스토리텔링 매체인 AR을 실제로 탐색하는 문헌을 활용하여 유적지에서의 AR 경험을 위한 프로토타입 디자인 프레임워크를 제안합니다. (이하 생략)

예시 논문 2번에 추천된 IPC 코드

IPC 코드 : G06F G 섹션 : 물리학 G06 : 전산; 계산 또는 계수 G06F : 전기에 의한 디지털 데이터처리

3.2.2. TF-IDF와 XGBClassifier를 이용하는 방법

특허문헌과 비특허문헌을 TF-IDF로 벡터화하기 위해서, 아나콘다를 설치한 후 가상환경을 만들어 파이썬 버전 3.6.13에서 Scikit-learn 라이브러리 문서전처리 기능의 TfidfVectorizer 클래스를 이용한다. 다음 단계로, 특허문헌들의 벡터값들을 독립변수로 하고 각 특허문헌의 메인 특허분류 코드를 종속변수로 하여 XGBClassifier 모델을 학습하기 위해서, 아나콘다를 설치한 후 가상환경을 만들어 파이썬 버전 3.6.13에서 xgboost 라이브러리의 XGBClassifier 클래스를 이용한다. 이후, 학습한 모델을 비특허문헌들의 벡터값들에 적용하여 각 비특허문헌에 특허분류 코드를 추천한다. 결과는 다음과 같다.

예시 논문 1번의 요약의 번역

에너지에 대한 대외 의존 문제와 국내/국제 환경 정책 방향은 재생 에너지 시스템에 대한 투자를 장려합니다. 신재생에너지시스템(RES)에서는 태양에너지를 기반으로 한 태양광발전(PV) 시스템이 매년 증가하고 있다. (이하 생략)

예시 논문 1번에 추천된 IPC 코드

IPC 코드 : G06F G 섹션 : 물리학 G06 : 전산; 계산 또는 계수 G06F : 전기에 의한 디지털 데이터처리

예시 논문 2번의 요약의 번역

유산에서 스마트 글래스 증강 현실(AR)의 적용이 증가하고 있음에도 불구하고 의미 있고 교육적인 몰입형 유산 경험을 설계하기 위한 기반이 될 수 있는 프레임워크는 없습니다. 이 기사에서는 정서적 경험을 학습과 연결하고 비교훈적인 스토리텔링 매체인 AR을 실제로 탐색하는 문헌을 활용하여 유적지에서의 AR 경험을 위한 프로토타입 디자인 프레임워크를 제안합니다. (이하 생략)

예시 논문 2번에 추천된 IPC 코드

IPC 코드 : A63F A 섹션 : 인간 필수품 A63 : 운동; 놀이; 오락; A63F : 카드게임, 보드게임 또는 룰렛게임; 작은 움직이는 물체를 사용하는 실내용게임; 비디오 게임; 그 밖에 분류되지 않는 게임

3.3. 소결론

특허문헌 텍스트 마이닝을 통해서 비특허문헌을 자동분류하기 위한 2가지 프로세스인, 텍스트 유사도 측정 알고리즘을 사용하는 방법과 텍스트 분류 알고리즘을 사용하는 방법을 'augmented reality'라는 키워드를 예시로 적용했다. 특허문헌은 미국의 등록특허 중 명칭, 요약, 대표청구항에 "augmented reality"와 정확히 일치하는 구문이 있는 문헌을 연구를 위해 예시로 들었고, 비특허문헌은 스코퍼스(SCOPUS)에 등록된 논문 중 제목, 요약문, 키워드에 "augmented reality"와 정확히 일치하는 구문이 있는 문헌을 연구를 위해 예시로 들었다. 연구는 모두 아나콘다를 설치한 후 가상환경을 만들어 파이썬 환경에서 진행되었다. 결과를 나타내기 위해서 모든 논문을 예시로 들 수 없으므로, 논문 중 2개('The role of virtual and augmented reality in occupational health and safety training of employees in PV power systems and evaluation with a sustainability perspective[48)]', 'A Design Framework for Smart Glass Augmented Reality Experiences in Heritage Sites[49)]')를 선택하여 결과를 작성했다.

먼저, 텍스트 유사도 측정 알고리즘을 사용하는 방법에서는 텍스트 유사도를 측정하려는 방법으로 TF-IDF와 SBERT를 이용하였는데, 두 가지 방법 모두 논문에 주요 키워드가 유사한 특허문헌을 나타내었다. 예시 논문 1번과 유사도가 1위인 특허는 TF-IDF와 SBERT 모두 같은 것

48) Begüm Erten et al., "The role of virtual and augmented reality in occupational health and safe ty training of employees in PV power systems and evaluation with a sustainability perspective", *Journal of Cleaner Production*, Vol.379 Part2(2022), Article No. 134499.

49) Mariza Dima, "A Design Framework for Smart Glass Augmented Reality Experiences in Heritage Sites", *Journal on Computing and Cultural Heritage*, Vol.15 No.4(2022), Article No. 66.

을 나타냈다. 또, TF-IDF에서 예시 논문 2번과 유사도가 2위인 특허와 SBERT에서 예시 논문 2번과 유사도가 1위인 특허가 같은 것을 나타냈다.

TF-IDF에서 예시 논문 1번의 요약과 유사도가 1위인 특허, 2위인 특허의 IPC 코드를 종합해서 섹션, 클래스, 서브 클래스까지만을 살펴보면 다음과 같다.

F21V : 조명장치 또는 그 시스템의 기능적 특징 또는 그 세부. 달리 분류되지 않는, 다른 물체와 조명장치의 구조적 결합
G02B : 광학요소, 광학계 또는 광학장치
G03H : 홀로그래픽 처리 또는 장치
G06F : 전기에 의한 디지털 데이터처리
G06K : 데이터의 인식; 데이터의 표시; 기록매체; 기록매체의 취급
G06T : 이미지 데이터 처리 또는 발생, 일반
G09G : 정적수단을 사용하여 가변정보를 표시하는 표시장치의 제어를 위한 장치 또는 회로
G10K : 음을 발생하는 장치; 소음 또는 기타 음향파를 방호하거나 감쇠시키는 방법 또는 장치 일반; 달리 분류되지 않는 음향
G21K : 달리 분류되지 않는 입자 또는 전리 방사 취급 기술; 조사장치; 감마선 또는 X선 현미경
H04N : 화상통신

예시 논문 1번과의 유사도가 1위인 특허의 IPC 코드는 'G06T, G09G, G06K, G02B'로 자동분류를 위한 추천에 도움이 될 것으로 보인다. 하지만, 예시 논문 1번과의 유사도가 2위인 특허의 경우, 요약을 예시 논문 1번의 요약과 비교했을 때도 유사하지 않고 IPC 코드 또한 자동분류를 위한 추천에 도움이 될 것으로 보이지 않는다. 예시 논문 2번의 요약과 유사도가 1위인 특허, 2위인 특허의 IPC 코드를 종합해서 섹션, 클래스, 서브 클래스까지만을 살펴보면 다음과 같다.

G01C : 거리, 수평, 방위의 측정; 측량; 항법; 자이로스코프 기기; 사진측량 또는 영상측량
G02B : 광학요소, 광학계 또는 광학장치
G06F : 전기에 의한 디지털 데이터처리
G06T : 이미지 데이터 처리 또는 발생, 일반
G06Q : 관리용, 상업용, 금융용, 경영용, 감독용 또는 예측용으로 특히 적합한 데이터 처리 시스템 또는 방법; 그 밖에 분류되지 않는 관리용, 상업용, 금융용, 경영용, 감독용 또는 예측용으로 특히 적합한 시스템 또는 방법
G06V : 이미지 또는 비디오 인식 또는 이해
H04N : 화상통신

예시 논문 2번과의 유사도가 1위인 특허, 2위인 특허의 IPC 코드는 모두 자동분류를 위한 추천에 도움이 될 것으로 보인다. TF-IDF와 코사인 유사도를 이용하는 방법의 결과를 전체적으로 확인했을 때, 유사도가 높은 특허들은 모두 자동분류에 도움을 받을 수 있을 것으로 보였으나, 본 연구에서의 연구 범위가 좁고 키워드가 한정되어 있으므로 몇몇 문헌들에는 도움을 받기 어려울 것으로 보였다. 이러한 한계를 극복한다면 TF-IDF와 코사인 유사도를 이용하는 방법은 비특허문헌을 자동분류하려는 방법으로 적절할 것으로 예상된다.

SBERT에서 예시 논문 1번의 요약과 유사도가 1위인 특허, 2위인 특허의 IPC 코드를 종합해서 섹션, 클래스, 서브 클래스까지만을 살펴보면 다음과 같다.

G02B : 광학요소, 광학계 또는 광학장치
G06K : 데이터의 인식; 데이터의 표시; 기록매체; 기록매체의 취급
G06T : 이미지 데이터 처리 또는 발생, 일반
G09G : 정적수단을 사용하여 가변정보를 표시하는 표시장치의 제어를 위한 장치 또는 회로

예시 논문 1번과의 유사도가 1위인 특허, 2위인 특허의 IPC 코드는 모두 자동분류를 위한 추

천에 도움이 될 것으로 보인다. 예시 논문 2번의 요약과 유사도가 1위인 특허, 2위인 특허의 IPC 코드를 종합해서 섹션, 클래스, 서브 클래스까지만을 살펴보면 다음과 같다.

> G02B : 광학요소, 광학계 또는 광학장치
> G06F : 전기에 의한 디지털 데이터처리
> G06K : 데이터의 인식; 데이터의 표시; 기록매체; 기록매체의 취급
> G06Q : 관리용, 상업용, 금융용, 경영용, 감독용 또는 예측용으로 특히 적합한 데이터 처리 시스템 또는 방법; 그 밖에 분류되지 않는 관리용, 상업용, 금융용, 경영용, 감독용 또는 예측용으로 특히 적합한 시스템 또는 방법
> G06T : 이미지 데이터 처리 또는 발생, 일반
> G06V : 이미지 또는 비디오 인식 또는 이해
> H04N : 화상통신

예시 논문 2번과의 유사도가 1위인 특허의 IPC 코드는 'G02B, G06F, G06T, G06Q, G06V'로 자동분류를 위한 추천에 도움이 될 것으로 보인다. 하지만, 예시 논문 2번과의 유사도가 2위인 특허의 경우, 요약을 예시 논문 2번의 요약과 비교했을 때 유사도가 1위인 특허에 비해 덜 유사하고 IPC 코드 또한 자동분류를 위한 추천에 도움이 될 것으로 보이지 않는다. SBERT와 코사인 유사도를 이용하는 방법의 결과를 전체적으로 확인했을 때, 유사도가 높은 특허들은 모두 자동분류에 도움을 받을 수 있을 것으로 보였으나, TF-IDF와 코사인 유사도를 이용하는 방법과 같은 한계가 나타났다.

다음으로, 텍스트 분류 알고리즘을 사용하는 방법에서는 텍스트 분류를 위한 방법으로 Logistic Regression과 XGBClassifier를 이용하였는데, 두 가지 방법 모두 논문에 핵심적인 IPC 코드를 추천해 주었다. Logistic Regression에서 예시 논문 1번에 추천된 IPC 코드는 'G06T'이고, 예시 논문 2번에 추천된 IPC 코드는 'G06F'이다. 두 IPC 코드 모두 'augmented reality'라는 키워드와 관계가 있으므로, 텍스트 분류 알고리즘이 자동분류를 위한 추천에 도움이 될 것으로 볼 수 있다. 하지만, TF-IDF와 Logistic Regression을 이용하는 방법의 결과를 전체적으로 확인했을 때, 대부분 논문에 'G02B, G06F, G06T, G06Q, G06V, H04N'을 추천하고 있으므로 효과가 없는 것으로 판단된다. 이러한 한계를 극복하기 위해서는 일반적인 텍스트 분류 알고리즘과 달리 모델을 학습하는 데에 종속변수로 사용된 IPC 코드를, 메인 IPC 코드뿐만 아니라 전체 IPC 코드를 학습하도록 하는 다변량 분류가 더욱 적절할 수 있다. 추가로, 추천해 주는 IPC 코드를 다양하게 하는 다중 레이블 분류도 효과적일 것이다. 그리고 텍스트 유사도 측정 알고리즘을 사용하는 방법에서 문제점을 해결하는 방법도 도움이 될 수 있다. XGBClassifier에서 예시 논문 1번에 추천된 IPC 코드는 'G06F'이고, 예시 논문 2번에 추천된 IPC 코드는 'A63F'이다. 예시 논문 1번에 추천된 IPC 코드의 결과로 TF-IDF와 Logistic Regression을 이용하는 방법도 TF-IDF와 XGBClassifier를 이용하는 방법과 같은 한계를 가지고 있는 것으로 보인다. 하지만 예시 논문 2번에 추천된 IPC 코드의 결과로, 학습에 많이 사용되지 않지만 유의미한 IPC 코드를 추천해주었다. TF-IDF와 XGBClassifier을 이용하는 방법의 결과를 전체적으로 확인했을 때도, TF-IDF와 Logistic Regression을 이용하는 방법에 비해 문제를 조금은 극복한 것으로 보이며, 특허문헌 텍스트 마이닝을 통해서 비특허문헌을 자동분류하기 위해서 텍스트 분류 알고리즘을 사용한다면 TF-IDF와 Logistic Regression을 이용하는 방법보다 TF-IDF와 XGBClassifier를 이용하는 방법이 효과적일 것으로 판단된다.

4. 결론

특허를 출원하기 위해서나 출원된 특허를 심사하기 위해 거쳐야 하는 선행기술조사를 위해서 비특허문헌에 대한 검색이 필요하다. 하지만 비특허문헌은 특허문헌과 달리 표준화되어 있지 않으며 통일된 검색체계를 제공하지 않으므로, 선행기술조사는 특허문헌 조사와 비특허문헌 조사로 나누어져야 하고 특히 비특허문헌 조사의 경우는 한 번에 진행하는 데에 어려움이 있다. 이를 극복하기 위해서, 본 연구에서는 비특허문헌에도 특허문헌 분류에 사용되는 분류방법을 적용하여, 특허문헌을 검색하는 것과 같은 방법으로 비특허문헌을 검색할 수 있는 시스템을 제안했다. 방법으로는 특허문헌에 특허분류 코드를 추천하거나 자동으로 분류해 주는 기계학습 방법을 비특허문헌에 적용하는 것이 효과적이라고 판단하여, 기계학습 알고리즘을 이용하여 비특허문헌인 논문에 특허문헌에 사용되는 IPC 코드를 자동분류하는 과정을 예시로 들고 이를 검토해 보았다. 크게 텍스트 유사도 측정 알고리즘을 사용하는 방법과 텍스트 분류 알고리즘을 사용하는 방법으로 나누어 확인해 보았으며, 텍스트 유사도 측정 알고리즘을 사용하는 방법은 TF-IDF와 코사인 유사도를 이용하는 방법과 SBERT와 코사인 유사도를 이용하는 방법을, 텍스트 분류 알고리즘을 사용하는 방법은 TF-IDF와 Logistic Regression을 이용하는 방법과 TF-IDF와 XGBClassifier를 이용하는 방법을 연구에 적용하였다.

연구의 결과를 살펴보았을 때, 텍스트 유사도 측정 알고리즘을 사용하는 방법은 TF-IDF와 코사인 유사도를 이용하는 방법과 SBERT와 코사인 유사도를 이용하는 방법 모두 자동분류에 도움을 받을 수 있을 것으로 보이며, 연구 환경에서의 한계를 극복한다면 효과적일 것으로 예측되었다. 하지만, 텍스트 유사도 측정 알고리즘을 사용하는 방법은 대상 특허문헌과 비특허문헌의 수가 많아질 수록 유사도를 측정하기 위해 들이는 시간이 늘어나는 것이 문제가 될 수 있다. 또한, 텍스트 분류 알고리즘을 사용하는 방법은 본 연구의 TF-IDF와 Logistic Regression을 이용하는 방법과 TF-IDF와 XGBClassifier를 이용하는 방법 모두 자동분류에 도움을 주기는 힘든 것으로 보인다. 하지만, TF-IDF와 Logistic Regression을 이용하는 방법보다는 TF-IDF와 XGBClassifier를 이용하는 방법이 효과적이며, 다른 방법론을 적용한다면 성능이 좋아질 것으로 예측한다.

전체적으로, 특허문헌 텍스트 마이닝을 통해서 비특허문헌을 자동분류는 앞으로 연구가 더욱 필요할 것이고, 텍스트 유사도 측정 알고리즘을 사용하는 방법과 텍스트 분류 알고리즘을 사용하는 방법에 관한 연구에 덧붙여 다른 텍스트 마이닝 알고리즘을 사용하는 효과적인 방법도 찾을 수 있을 것이다. 이를 위해서 본 연구에서의 한계와 문제점을 정리해 보면 다음과 같다. 첫째, 다른 자동분류 알고리즘과 달리 정확도를 객관적으로 판단하기 어렵다는 점이다. 특허를 분류하는 과정 역시 특허청 특허분류부여 전문기관인 한국특허기술진흥원이 국내에 출원된 특허·실용신안 및 PCT(Patent Cooperation Treaty, 특허협력조약) 출원에 대해 기술 내용을 파악하여 해당 기술 분야별로 적절한 특허분류를 부여하지만[50][51], 비특허문헌의 경우 이를 정확하게 판단하기 어렵고 사람이 일일이 결과를 확인해 보아야 한다. 둘째, 연구 범위가 넓을수록 좋은 성능을 가진다는 점이다. 기계학습 알고리즘의 경우 학습에 사용된 데이터의 품질과 양이 핵심이므로[52] 이를 준비하는 과정에 큰 노력을 기울여야 할 것이다. 마지막으로, 텍스트 분류 알고리즘을 사용하는 방법을 위해서는 다변량 분류와 다중 레이블 분류를 모두 적용한 분류 방

50) 한국특허기술진흥원, “특허분류부여”, 한국특허기술진흥원, <https://www.kipro.or.kr/business/patentClassification>, 검색일: 2024. 1. 26.

51) 특허법 제58조 및 실용신안법 제15조.

52) 서대호, “AI에 절대적인 데이터…거래 활성화 플랫폼 필요”, ZDNET Korea, <https://zdnet.co.kr/view/?no=20220416144501>, 검색일: 2024. 1. 26.

법이 필요하다는 점이다. 특허문헌은 하나의 특허에 여러 개의 특허분류를 가질 수 있으므로, 학습 데이터에 여러 개의 특허분류를 적용하기 위한 다변량 분류가 필요하다. 비특허문헌도 마찬가지로 여러 개의 특허분류를 가질 수 있어야 하므로 다중 레이블 분류가 적용되어야 올바른 특허문헌을 검색하는 것과 같은 방법으로 비특허문헌을 검색할 수 있는 시스템이 구축될 수 있을 것이다.

참고문헌

단행본(국내 및 동양)

수다르산 라비찬디란, 「구글 BERT의 정석」, 한빛미디어, 2022.

안상준 외 1인, 「딥 러닝을 이용한 자연어 처리 입문 2권」, 위키독스, 2023.

안상준 외 1인, 「딥 러닝을 이용한 자연어 처리 입문 3권」, 위키독스, 2023.

전창욱 외 3인, 「텐서플로 2와 머신러닝으로 시작하는 자연어 처리」, 위키북스, 2022.

코리 웨이드, 「XGBoost와 사이킷런을 활용한 그레이디언트 부스팅」, 한빛미디어, 2022.

학술지(국내 및 동양)

김판준, "자질선정을 통한 국내 학술지 논문의 자동분류에 관한 연구", 「정보관리학회지」, 제39권 제1호(2022).

김현종 외 3인, "비지도학습 기반의 행정부서별 신문기사 자동분류 연구", 「한국산학기술학회논문지」, 제21권 제9호(2020).

박영규, "선택발명의 신규성·진보성 판단을 위한 선행기술의 인정범위", 「지식재산연구」, 제14권 제4호(2019).

박진우 외 4인, "한국어 특허 문장 기반 CPC 자동분류 연구 ―인공지능 언어모델 KorPatBERT를 활용한 딥러닝 기법 접근", 「지식재산연구」, 제17권 제3호(2022).

박찬정 외 3인, "용어 클러스터링을 이용한 특허문서 자동 IPC 분류", 「한국정보기술학회논문지」, 제12권 제9호(2014).

오석범/강현국, "머신러닝 알고리즘을 이용한 표준문서의 ICS 코드 자동분류(ICAC)", 「표준인증안전학회지」, 제11권 제2호(2021).

임소라/권용진, "특허문서 필드의 기능적 특성을 활용한 IPC 다중 레이블 분류", 「인터넷정보학회논문지」, 제18권 제1호(2017).

정연경, "학문분류표의 재설정에 관한 연구", 「정보관리학회지」, 제17권 제2호(2000).

주시형, "특허인용의 등록유지 기간에의 영향 - 한국특허 인용 정보를 활용한 분석", 「지식재산연구」, 제15권 제4호(2020).

학술지(서양)

Ashish Vaswani et al., "Attention Is All You Need", *arxiv*, (2017).

Begum Erten et al., "The role of virtual and augmented reality in occupational health and safety training of employees in PV power systems and evaluation with a sustainability perspective", *Journal of Cleaner Production*, Vol.379 Part2(2022)

David R. Cox, "The regression analysis of binary sequences (with discussion)", *J Roy Stat Soc B*, Vol.20(1958).

Gema Velayos-Ortega & Rosana Lopez-Carreno, "Non-Patent Literature", *Encyclopedia*, Vol.1 No.1(2021).

Jacob Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *arxiv*, (2018).

Mariza Dima, "A Design Framework for Smart Glass Augmented Reality Experiences in Heritage Sites", *Journal on Computing and Cultural Heritage*, Vol.15 No.4(2022).

Nils Reimers & Iryna Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks", *arxiv*, (2019).

Singhal Amit, "Modern Information Retrieval: A Brief Overview", *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, Vol.24 No.4(2001).

Tianqi Chen & Carlos Guestrin, "Xgboost: A scalable tree boosting system", *arxiv*, (2016).

인터넷 자료

김주동 외 2인, "'스마트한 텍스트 분석을 향한 첫걸음, KoreALBERT", SAMSUNG SDS, <https://www.samsungsds.com/kr/insights/techtoolkit_2021_korealbert.html>, 검색일: 2024. 1. 26.

서대호, "AI에 절대적인 데이터…거래 활성화 플랫폼 필요", ZDNET Korea, <https://zdnet.co.kr/view/?no=20220416144501>, 검색일: 2024. 1. 26.

특허청, "기술-품목-특허 연계표", 특허청, <https://www.kipo.go.kr/ko/kpoContentView.do?menuCd=SCD0200273>, 검색일: 2024. 1. 26.

특허청, "CPC 및 IPC 분류코드", 특허청, <https://www.kipo.go.kr/ko/kpoContentView.do?menuCd=SCD0200269>, 검색일: 2024. 1. 26.

한국연구재단, "연구분야분류표", 한국연구재단, <https://www.nrf.re.kr/biz/doc/class/view?menu_no=322>, 검색일: 2024. 1. 26.

한국특허기술진흥원, "선행기술조사", 한국특허기술진흥원, <https://www.kipro.or.kr/business/priorArtSearch>, 검색일: 2024. 1. 26.

한국특허기술진흥원, "특허분류부여", 한국특허기술진흥원, <https://www.kipro.or.kr/business/patentClassification>, 검색일: 2024. 1. 26.

한국특허기술진흥원, "특허분류 연계정보", 한국특허기술진흥원, <https://cls.kipro.or.kr/classification/linkedTable/search>, 검색일: 2024. 1. 26.

Jiaming Yuan, "XGBoost", DMLC XGBOOST, <https://xgboost.readthedocs.io/en/stable/index.html>, 검색일: 2024. 1. 26.

Jiaming Yuan, "xgboost", GitHub, <https://github.com/dmlc/xgboost>, 검색일: 2024. 1. 26.

Wikipedia, "INID", Wikipedia, <https://en.wikipedia.org/wiki/INID>, 검색일: 2024. 1. 26.

연구보고서

장경선 외 5인, "해외 지식재산권 데이터 관리현황 조사 및 연구", 특허청, 2006.

기타자료

권오진 외 4인, "핵심정보자원 연계를 통한 국내 특허 인용 정보 생성 방법에 관한 연구", 한국기술혁신학회 학술대회, 한국기술혁신학회, 2005.

심우철 외 4인, "한국 특허문헌 특성 및 딥러닝 기반 분류모델을 고려한 CPC 자동분류에 관한 연구", 한국정보과학회 학술발표논문집, 한국정보과학회, 2020.