Journal of Korean Library and Information Science Society
Vol. 53, No. 2

AttentionMesh를 활용한 국가과학기술표준분류체계 소분류 키워드 자동추천에 관한 연구*

A Study on Automatic Recommendation of Keywords for Sub-Classification of National Science and Technology Standard Classification System Using AttentionMesh

박 진 호(Jin Ho Park)**
송 민 선(Min Sun Song)***

〈 목 차 〉

요 약 : 이 연구의 목적은 국가과학기술표준분류체계의 소분류 용어를 기계학습 알고리즘을 적용하여 기술키워드 변환하는 것이 목적이다. 이를 위해 본 연구에서는 주제어 추천에 적합한 학습 알고리즘으로 AttentionMeSH를 활용했다. 원천데이터는 한국과학기술기획평가원이 정제한 2017년부터 2020년까지 4개년 연구현황 파일을 사용하였다. 학습은 과제명, 연구목표, 연구내용, 기대효과와 같이 연구내용을 잘 표현하고 있는 4개 속성을 사용했다. 그 결과 임계치(threshold)가 0.5일 때 MiF 0.6377이라는 결과가 도출됨을 확인하였다. 향후 실제 업무에 기계학습을 활용하고, 기술키워드 확보를 위해서는 용어관리체계 구축과 다양한 속성들의 데이터 확보가 필요할 것으로 보인다.

주제어 : 국가과학기술표준분류체계, 주제어 추천, 학습 알고리즘, 주제어 학습, AttentionMeSH

ABSTRACT : The purpose of this study is to transform the sub-categorization terms of the National Science and Technology Standards Classification System into technical keywords by applying a machine learning algorithm. For this purpose, AttentionMeSH was used as a learning algorithm suitable for topic word recommendation. For source data, four-year research status files from 2017 to 2020, refined by the Korea Institute of Science and Technology Planning and Evaluation, were used. For learning, four attributes that well express the research content were used: task name, research goal, research abstract, and expected effect. As a result, it was confirmed that the result of MiF 0.6377 was derived when the threshold was 0.5. In order to utilize machine learning in actual work in the future and to secure technical keywords, it is expected that it will be necessary to establish a term management system and secure data of various attributes.

KEYWORDS : National Science and Technology Standard Classification System, Keyword Recommendation, Learning Machine Algorithm, Keyword Learning, AttentionMeSH

---

* 이 논문은 2021년 한국과학기술기획평가원의『국가과학기술표준분류 소분류 기술키워드화 운영 방안 설계』 일부분을 수정・보완한 것임.

** 한성대학교 크리에이티브 인문학부 도서관정보문화트랙 조교수
(jhp@hansung.ac.kr / ISNI 0000 0004 7641 0372) (제1저자)
*** 대림대학교 도서관미디어정보과 조교수(songsery@daelim.ac.kr / ISNI 0000 0004 9246 0812) (교신저자)

• 논문접수: 2022년 5월 25일 • 최초심사: 2022년 6월 4일 • 게재확정: 2022년 6월 21일
• 한국도서관・정보학회지, 53(2), 95-115, 2022. http://dx.doi.org/10.16981/kliss.53.2.202206.95

Ⅰ. 서 론

국가과학기술표준분류체계(이하, 표준분류체계)는 과학기술분야 연구개발사업의 관리를 목적으로 제정되었다. 표준분류체계는 「과학기술기본법. 법률 제18069호」에 근거하는데 동 법의 27조 1항에서는 이 분류체계의 목적이 과학기술 관련 정보・인력・연구개발사업 등을 효율적으로 관리하기 위함으로 명시되어 있다. 동법 시행령 41조 7항에는 표준분류체계의 활용 방법 세 가지를 구체적으로 명시하고 있다. 첫 번째는 국가연구개발사업의 연구기획・평가 및 관리이고, 두 번째는 과학기술예측 및 기술수준평가이며, 마지막 세 번째는 과학기술지식・정보의 관리・유통이다. 시행령에서 제시하고 있는 활용의 핵심은 '평가'와 '관리'이다. 단, 여기서 말하는 '평가'와 '관리'는 특정 연구의 품질유지를 의미하는 것이 아니며, 표준분류체계를 바탕으로 한 해 동안 정부가 투자해 추진된 국가연구개발사업의 현황을 알아보고 그 결과를 확인하고자 하는 절차이다. 이 업무는 한국과학기술기획평가원이 담당하고 있으며, 구체적인 결과물은 『국가연구개발사업조사분석보고서』와 『국가연구개발사업조사분석데이터』의 형태로 산출된다. 최종 결과보고서와 통계자료는 정부 및 민간 영역의 기술 분야별 투자 현황과 같은 R&D 기초 통계자료 분류, 정부 R&D 사업의 예산 조정 및 배분, 과학기술 분야 국가연구개발사업 과제 선정 및 평가・관리 등의 영역에서 활용된다(한국과학기술기획평가원, 2019).

이 표준분류체계는 2002년 처음 제정된 후 지속적인 수정・보완 작업을 거쳐왔다. 차기 표준분류체계 개정은 2022년 중 예정되어 있으며, 이번 개정의 핵심 개선 방향은 분류 구조를 단순화하고, 신기술 분야에 대한 신속한 반영을 통해 유관 분류체계와 연계를 강화함으로써 활용도를 높이고자 하는 데 있다. 이를 실현하기 위한 구체적인 방안은 현 '4계층 구조(연구분야 〉 대 〉 중 〉 소)'를 '3계층 구조(연구분야 〉 대 〉 중)'로 단순화하고 기존 소분류는 기술키워드(technical keywords)로 대체해 활용하는 것이다(과학기술정보통신부, 2019). 여기서 기술키워드란 해당 연구의 주제어를 의미한다. 이번 개정에서 분류 구조를 단순화하는 주요 이유는 5년 주기로 표준분류체계를 개정하는 과정에서 가장 하위에 속하는 소분류 체계가 전문가들의 검토를 거쳐 대, 중 분류의 하위에 속하는 주제 분야로 분류되면서 개정 주기 동안 등장한 새로운 과학기술 연구의 반영이 어렵고 분류체계에 명확하게 속하지 않는 주제의 경우는 반영이 안 되는 문제가 지속적으로 발생해왔기 때문이다. 또한 다양한 주제 분야에서 공통적으로 연구가 추진되는 분류의 경우도 기존 표준분류체계로는 반영이 어려운 문제가 발생한다. 소분류의 기술키워드화는 이런 문제점들을 해결하고자 하는 방안으로 볼 수 있다.

현재 4계층 분류체계 구조의 소분류를 기술키워드화 체계로 전환하면 개별 연구자들은 자유롭게 자신의 연구를 대표할 수 있는 용어를 직접 입력하고 활용할 수 있다. 하지만 3계층 구조와 소분류 기술키워드화를 통해 표준분류체계의 유연성을 확보하더라도 운영 주체인 한국과학기술기획평가원

은 기술키워드에 새로운 운영체계를 갖출 필요가 있다. 즉 기술키워드화가 되더라도 현재처럼 연구자들이 입력한 기술키워드를 표준화하여 운영・관리해야 한다. 또한 연구자들의 경우도 본인이 입력한 기술키워드가 적합한 것인지에 대한 검증과 함께 본인 연구 분야에 적절한 키워드를 도출하기 위한 별도의 노력이 필요하다. 즉, 기술키워드화를 한다고 하더라도 운영 주체는 표준화된 절차에 따른 용어 관리가 필요하며, 다른 연구에서 사용하고 있거나 신규 표준 용어로 등록된 용어에 대한 활용 가능 여부를 지속적으로 확인해야 한다. 그래야 현재의 표준분류체계 소분류를 기술키워드로 전환하여 새로운 연구 흐름을 반영하고, 학제 간 연구 현황 분석 등에 분류체계를 활용하고자 하는 목적을 달성할 수 있을 것이다.

본 연구는 국가과학기술표준분류체계의 소분류를 기술키워드화하고, 분류체계에 활용할 수 있는 키워드 도출 방법의 하나로 학습 알고리즘에 의한 자동추천 방식을 적용해보고 그 결과를 분석하는 것이 목적이다. 자동추천 방식의 활용은 기존 소분류 체계에서의 용어가 갖는 문제점을 개선하기 위해서는 아니며, 향후 변경되는 운영방식을 보완할 수 있는 방안 중 기존 용어를 활용하면서 초기 양적 충당을 달성하기에 적합하기 때문이다. 이는 운영 주체인 한국과학기술기획평가원에서 도입할 수 있는 방법 중 하나로, 개정되는 분류체계 운영 초기에 발생할 수 있는 기술키워드의 양적 충당과 개별 연구자들이 표준용어를 선택하는 데 겪을 수 있는 어려움 해결을 기대할 수 있는 방법 중 하나이다. 이를 위해 본 연구에서는 주제어 추천에 적합한 학습 알고리즘을 선택하여 기존 연구현황 데이터세트를 학습 및 결과 확인 원천데이터로 활용하여 자동추천을 시도하였다. 학습 알고리즘은 AttentionMeSH를 활용했고, 원천데이터는 한국과학기술기획평가원이 정제한 2017년부터 2020년까지 4개년 연구현황 파일 데이터를 대상으로 했다.

Ⅱ. 이론적 배경과 선행연구

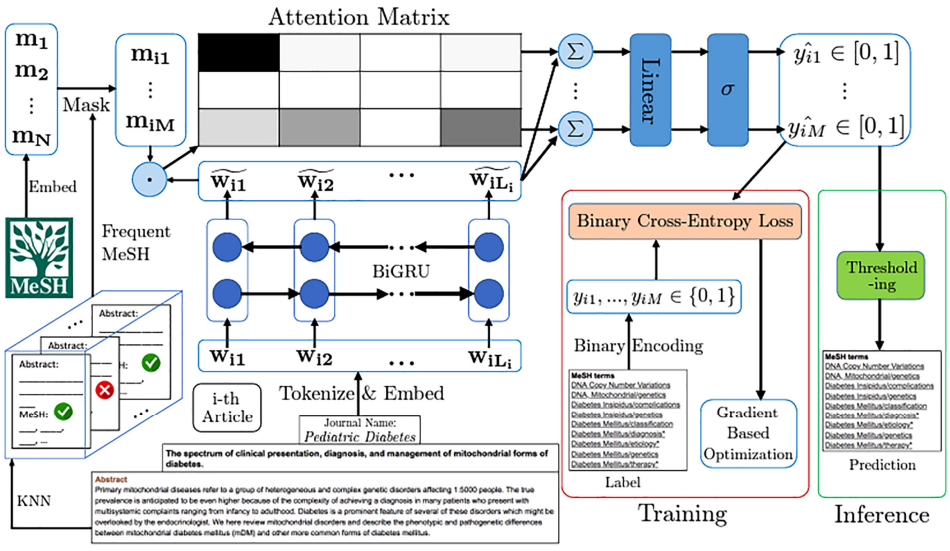
1. AttentionMeSH

본 연구에서 활용한 학습 알고리즘은 AttentionMeSH이다. 이름에서 유추할 수 있듯이 이 알고리즘은 미국국립의학도서관(National Library of Medicine, NLM)의 주제명을 대상으로 개발되었다. MeSH는 의학주제명표목(Medical Subject Headings)의 줄임말로 NLM이 운영하는 의학분야의 주제명표목이다. MeSH는 1960년에 처음 소개된 이후 지속적으로 개정되어 새로운 개념, 용어를 추가하고 있으며 전 세계 의학 분야의 표준 통제어휘집으로 활용되고 있다. NLM은 공식적으로 Medical Text Indexer(이하 MTI)를 활용하여 수동으로 주제어를 입력하는 방식을 개선하고자 노력해 왔다. 이런 노력의 일환으로 NLM은 자체적인 색인어 도구 개발 외에도 외부 전문가

의 참여가 가능한 대회를 추진하였는데 BioASQ Challenge가 대표적인 사례이다.

BioASQ는 생물의학분야의 의미 색인과 질의응답 색인, 질의응답 데이터세트 구성을 목적으로 하는 대회로 미국 국립보건원(National Institutes of Health, NIH)과 NLM의 지원 하에 운영된다. 이 대회에서 자동색인, 주제어 추천과 관련 있는 대표적인 모델은 두 가지로, 하나는 DeepMeSH이며, 다른 하나는 AttentionMeSH이다. DeepMeSH는 BioASQ Task5a 챌린지에서 우승했던 색인 모델이다(Jin et al., 2018). BioASQ Task5a는 2017년에 개최되었던 대회로 PubMed가 학술지의 초록을 색인화하는 프로세스를 기준으로 한다. 당시 DeepMeSH 학습 데이터는 PubMed의 큐레이터가 MeSH 용어를 할당한 2013년부터 2017년까지의 기사자료 데이터를 활용하였다. 한편 AttentionMeSH는 2018년에 있었던 BioASQ Task6a의 결과물이다.

2018년에는 6a와 6b 두 개의 과제가 동시에 진행되었는데, 6a는 생의학 분야의 대규모 온라인 의미적 색인(large-scale biomedical semantic indexing)이며, 6b는 생의학 분야 의미적 질의응답(biomedical semantic question answering) 과제였다. AttentionMeSH는 미국 버팔로 뉴욕주립대학(The State University of New York at Buffalo, USA) 연구진의 결과물로써(BioASQ, n.d. -a, BioASQ, n.d. -b). 학습데이터와 알고리즘이 논문 등을 통해 공개되어 있어 본 연구에서 학습 알고리즘으로 활용했다. Jin et al.(2018)은 AttentionMeSH가 갖는 가장 큰 특징으로 모델의 단순성과 해석의 용이함을 들었다. AttentionMeSH는 BioASQ Chanllenge Task6a의 마지막 주에 미완성 모델을 사용해 정확도(MiF, Micro F-measure) 0.6635를 달성했으며, 이 후 완성된 모델로 0.684에 가까운 MiF 성능을 달성했다. AttentionMeSH의 구성은 〈그림 1〉과 같다.



〈그림 1〉 AttentionMeSH 구성도(Jin et al., 2018)

〈그림 1〉에서 보듯이 AttentionMeSH 알고리즘에서는 문서 내의 초록, 제목, 저널 제목 순서로 문서의 단어가 입력된다. 이 입력값들은 양방향 순환신경망(BiGRU)에 공급되어 문맥을 인식한다. 이후 학습용 말뭉치로부터 k-NN(k-nearest neighbors) 알고리즘을 활용하여 자주 활용되는 MeSH 용어가 후보로 추출되며, MeSH가 포함되어 있으면 해당 용어만 attention 메커니즘에 포함되고, 이를 마스킹 매커니즘(masking mechanism)이라 부른다. MeSH와 관련된 특정 문서 표현과 연결되는 각각의 후보 MeSH 용어에는 attention 가중치를 할당한다.

AttentionMeSH가 주제어 추천을 위해 처리하는 각 단계와 처리방식은 〈표 1〉과 같이 간단히 정리할 수 있다.

〈표 1〉 AttentionMeSH의 주제어 추천 처리 단계와 방식(Jin et al., 2018)

처리 단계	처리 방식과 내용
Document Representation	• 색인 생성을 위해 각 기사의 저널 이름, 제목, 초록의 용어를 토큰화하는 단계 • $E \in R^{\lvert V \rvert \times d_{e1}}$ (|V|는 어휘크기, d_{e1}은 임베딩 크기) • $D = [W_1 \dots W_L]^T \in R^{L \times d_{e1}}$ (L은 저널 이름, 제목 및 초록에 있는 단어의 수, w_i는 위치 i에 있는 단어에 대한 벡터) • $\widetilde{D} = BiGRU(D) = [\widetilde{W}1 \dots \widetilde{W}_L]^T \in R^{L \times 2d_h}$ ($\widetilde{w_i}$ 는 각 단어의 연결된 순방향 및 역방향 은닉 상태에 해당하고 d_h는 BiGRU의 은닉 크기)
MeSH Representation and Masking	• MeSH 임베딩 행렬 '$H \in R^{N \times d_{e2}}$'를 학습('N'은 모든 MeSH 항의 수, 'd_{e2}'는 임베딩 크기) • k-NN 전략을 사용하여 각 기사에 대해 훈련할 특정 MeSH 용어 하위 집합을 선택 • 각 초록은 단어 벡터의 IDF 가중 합 • $d = \frac{\sum_{i=1}^{n} IDF_i \times W_i}{\sum_{i=1}^{n} IDF_i} \in R^{d_{e1}}$ (W_i는 해당 단어의 벡터, IDF_i는 이 단어의 역문헌 빈도) • 초록들 간 표현의 코사인 유사도를 계산함 • $Similarity(i,j) = \frac{\overset{T}{d}_i d_j}{\lVert d_i \rVert \times \lVert d_j \rVert}$ (각 기사에 대해 코사인 유사도를 기반으로 K개의 최근접 이웃 탐색 및 MeSH 용어 빈도 계산) • $H^l = [m_1 m_2 \dots m_M] \in R^{M \times d_{e2}}$ (가장 빈번한 MeSH 용어는 각 기사에 대응하여 학습되고 마스킹 된 'MeSH 임베딩 H^l'로 표시)
Attention Mechanism	• 문서 표현과 마스킹 된 MeSH 표현을 획득한 후 유사성을 계산함 • $S = H^l \widetilde{D}^T = [\widetilde{D}m_1 \dots \widetilde{D}m_M]^T \in R^{M \times L}$ • SoftMax 함수를 사용하여 단어 축에 대한 정규화를 거쳐 각 MeSH 용어에 대한 가중치 속성 획득 • SoftMax(Sim) $= [SoftMax(\widetilde{D}m_1) \dots SoftMax(\widetilde{D}m_M)]^T = [\alpha_1 \dots \alpha_M]^T \in [0,1]^{M \times L}$ • '$\alpha_j \in [0,1]^L$'는 attention 가중치로 MeSH용어와 '$\sum_{k=1}^{L} \alpha_{jk} = 1$' 일 경우에 대한 가중치

처리 단계	처리 방식과 내용
Classification	• 각 MeSH 용어에 대해 attention 가중치 • $R_j = \alpha_j \widetilde{D} \in R^{2d_h}$ (R_j는 특정 문서를 표현하는 MeSH 용어 j) • $\hat{y}_j = \sigma(\overset{T}{R}_j \overset{'}{m}_j + b_j) \in [0,1]$ ($\overset{'}{m}_j$과 b_j는 MeSH 용어 j에 대해 학습 가능한 선형 투사 매개변수) • 최종 모델은 P(MeSH j indexed | Journal, Title, Abstract) = $\hat{y}_j$
Training	• 모델링한 조건부 확률을 획득한 후 각 MeSH 용어에 대한 이진 교차 엔트리 손실을 계산함 • $L_j = -(y_j \log(\hat{y}_j) + (1-y_j)\log(1-\hat{y}_j))$ ($y_j \in 0,1$ 는 MeSH j의 실측 레이블, $y_j=0$은 MeSH j가 사람에 의해 주석 처리되지 않았음을 의미하고, $y_j=1$은 MeSH j가 주석 처리되었음을 의미) • 총 손실은 다음과 같이 계산함 • $L = \frac{1}{M}\sum_{j=1}^{M} L_j$
Inference	• 추론 시 예측확률이 조정된 임계값(threshold)보다 큰 MeSH 용어를 채택 • (predict MeSH j) = $\| (\hat{y}_j > p_j)$ (p_j는 MeSH 용어 j에 대한 조정된 임계값, 임계값은 MiF를 최대화하도록 조정) • $p_1, ..., p_N = \underset{p_1, ..., p_N}{\arg max}\, MiF(Model, p_1, ..., p_N)$

〈표 1〉의 단계 후 최종 결과 평가는 Micro-F 값으로 결정하게 되는데, Micro-F는 MiP(미세 정확률, micro-precision), MiR(미세 재현율, micro-recall)을 조합한 형태로 다음과 같이 계산한다.

$$Micro - F = \frac{2 \cdot MiP \cdot MiR}{MiP + MiR}$$

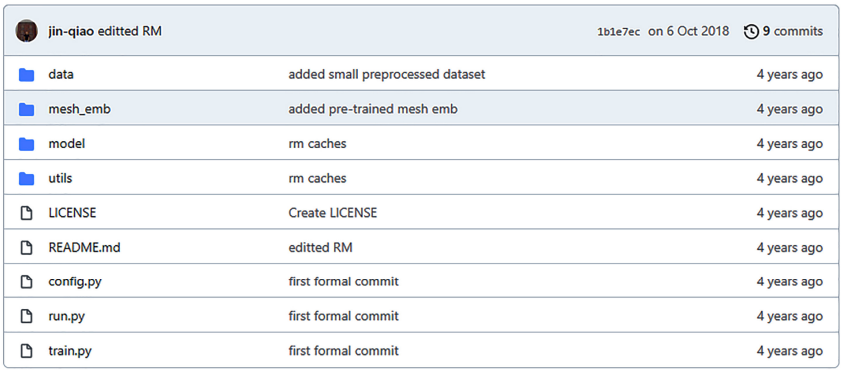
미세 정확률(MiP)과 미세 재현율(MiF)를 산정하는 방식은 다음과 같다.

$$MiP = \frac{\sum_{i=1}^{Na}\sum_{j=1}^{N} y_{ij} \cdot \hat{y}_{ij}}{\sum_{i=1}^{Na}\sum_{j=1}^{N} \hat{y_{ij}}}$$

$$MiR = \frac{\sum_{i=1}^{Na}\sum_{j=1}^{N} y_{ij} \cdot \hat{y}_{ij}}{\sum_{i=1}^{Na}\sum_{j=1}^{N} y_{ij}}$$

위 식에서 i는 기사에 대한 색인, j는 MeSH 용어에 대한 색인이다. Na는 테스트 세트의 기사 수이고, N은 모든 MeSH 용어의 수를 나타낸다. y_{ij} 및 $\hat{y_{ij}}$는 모두 MeSH 용어 j가 기사 i에 포함되어 있는지 여부를 나타내는 이진 인코딩 변수(binary encoded variables)를 의미한다.

본 연구는 주제어 학습과 추천을 위한 알고리즘 개발이 목적이 아니라 특정 알고리즘을 활용해 주제어 추천이 가능한지와 그 결과로 향후 해결방안을 모색하기 위한 것으로 이상에서 설명한 AttentionMeSH 알고리즘을 그대로 활용했다. 해당 알고리즘은 Github에 등록된 오픈 소스를 활용했다(〈그림 2〉 참조).



(출처: https://github.com/Andy-jqa/AttnMeSH)

〈그림 2〉 AttentionMeSH Github 페이지

본 논문에서 국가과학기술표준분류체계 관련 용어처리와 최종 결과 확인은 〈표 1〉의 절차를 거쳐 도출된 MiF 값을 통해 확인했다. 알고리즘은 수정하지 않고 그대로 활용했는데, 〈표 1〉에서 AttentionMeSH가 토큰화 기준으로 삼았던 저널명, 기사제목, 초록은 국가과학기술표준분류체계의 원 소스에 적합한 요소인 과제명, 연구목표, 연구내용, 기대효과로 변경 적용했다.

2. 선행연구

본 연구의 내용과 관련된 선행연구는 주제어 자동 추천 관련 연구, 국가과학기술표준분류체계를 대상으로 한 연구, 기계학습을 활용한 연구로 구분할 수 있다.

문헌정보학 분야에서의 주제어 자동 추천 관련 연구는 이용자 개인화 서비스와 정보검색 분야에서 주로 이루어졌는데, 대부분의 연구들에서 로그 데이터 분석 방법을 적용했다.

이소영, 정영미(2006)는 국내 포털에서 12일간 수집한 이용자 로그 데이터를 활용한 개인 검색 서비스 모형 설계 연구를 수행하고, 실험 결과에 대한 이용자 만족도를 리커드 5점 척도로 조사하여 높은 수준의 만족도를 확인했다.

김광영, 곽승진(2010) 역시 과학기술학회마을에서 제공하는 검색서비스를 분석해 개인화 검색 서비스를 개발했다. 이 연구는 기존 과학기술학회마을 검색서비스가 개인의 성향 정보를 반영하지

못한다는 점을 지적하고, 질의어 기반의 개인화 검색시스템, 논문과 관련된 공동 저자 내비게이션 시스템, 저자키워드 기반 주제어 추천 시스템과 유사한 사용자 자동 추천 시스템을 개발・구현하고 시스템 이용 만족도 조사에서 일반 검색보다는 개인화 검색 결과의 만족도가 높음을 확인했다.

조현양(2017; 2020)은 특정 이용자 집단을 대상으로 서비스를 제공하는 기관들의 추천 시스템에 대한 연구를 주로 수행하였는데, 2017년 연구에서는 국립어린이청소년도서관에서 제공하는 501권의 유아 및 아동도서를 대상으로 책소개 정보를 활용하여 개인별 성격유형에 적합한 도서를 추천하는 서평 자동분류시스템을 개발했다. 2020년 연구에서는 국립장애인도서관의 이용자 데이터 및 이용 내역 데이터를 기반으로 이용자의 선호 주제 분야 및 관심 키워드를 분석하고, 이용자 관심 정보를 반영할 수 있는 도서 자동 분류 엔진을 설계하고 이용자 선호도 기반 도서추천시스템을 제안한 바 있다.

주제분류 관련해서는 김광영, 곽승진(2011)이 이용자 성향에 맞는 정확한 검색결과 도출을 위해 주제 분류와 하이브리드 기반의 이용자 프로파일을 구성・적용한 연구를 수행하였다. 이 연구에서는 제안한 개인화 검색시스템의 성능 평가를 위해 과학기술학회마을 논문 80여만건을 이용해 전문가들이 일반 키워드와 중의성을 가진 키워드들의 이용 직접 적합성을 평가했다. 그 결과 전문가가 직접 "컴퓨터공학"과 "문헌정보학" 분야에서 평가한 국내과학기술논문 결과에서도 제안한 개인화 검색시스템이 일반 검색시스템보다 정확도가 더 높음을 확인했다.

한편, 국가과학기술표준분류체계와 관련해서는 한희준, 최윤수, 최성필(2018)이 이용자의 정보서비스 이용행태를 분석하여 검색 의도와 관심 분야를 국가과학기술표준분류체계를 기반으로 파악해 개인화 서비스에 적용하는 연구를 진행했다. 이 연구에서는 실시간 관심분야 추적, 관심 태그 클라우드 제공, 관심 분야 기반 추천정보 제공, 검색 결과 개인화 네 가지 기능으로 구성된 과학기술정보 개인화 서비스를 개발하여 전문가 실험집단과 통제집단과의 검색 성능 비교를 통해 개인화 정보의 적합성 및 개인화 기능 유용성을 평가했다. 그 결과 연구에서 제안된 개인화 서비스가 비교 대상 서비스보다 검색 성능이 더 우수한 것으로 나타났으며 더 높은 유용성을 제공하는 것을 입증했다.

최종윤, 한혁, 정유철(2020)은 국가과학기술표준분류체계를 기반으로 연구보고서의 자동분류에 대한 연구를 진행한 바 있다. 이 연구에서는 연구자들이 국가과학기술정보서비스(NTIS)에 보고서를 제출할 때 수동으로 분류코드를 입력하는 과정에서 약 2,000여 개의 분류체계 중 적합한 코드 선택이 어렵다는 문제를 해결하기 위한 자동분류 방식을 제시하고자 하였다. 자동분류 기법 도출을 위해 한국과학기술정보연구원(KISTI)이 보유하고 있는 5년간(2013년~2017년)의 연구보고서 메타정보를 활용해 국가과학기술표준분류체계 중 중분류체계 210여 개를 선별하고, 가장 영향력 있는 필드인 과제명(제목)과 키워드만을 이용한 TK_CNN 기반의 딥러닝 기법을 제안하였다. 제안 모델을 활용한 실험 결과, 기존의 기계학습법들(예, Linear SVC, CNN, GRU 등)과 비교하여 Top-3 F1 점수 기준으로 1~7%에 이르는 성능 우위를 확인했다.

김윤정, 신동구, 정회경(2021)은 기계학습을 활용해 R&D 과제의 연구 분야 추천 서비스 방안을 제시하였다. 이 연구는 기계학습을 이용해서 국가 R&D 과제 데이터를 학습하고 과제의 연구분야 소분류를 자동분류하여 연구자의 편의성을 높이기 위한 목적 하에 수행되었다. 연구 과정의 학습과 검증에 활용한 원천데이터는 국가연구개발사업 조사 분석을 통해 확정된 국가 R&D 과제로 2013년부터 2020년까지 약 45만건을 대상으로 하였다. 연구 결과, 제안한 모델 활용 실험에서 과제정보 필수항목으로 사용되는 소분류의 정확도가 90.11%로 도출됨을 확인했다.

송민선, 박진호(2021)는 현재의 국가과학기술표준분류의 소분류체계 기술키워드화를 위한 방법 중 하나로 표준기반의 용어관리체계를 제안했다. 여기서 대상 표준은 SKOS(Simple Knowledge Organization System)와 ISO/IEC 11179이다. SKOS를 대상으로 한 이유는 해외의 다양한 과학기술관련 용어관리체계를 조사한 바 대부분 용어관리로 SKOS를 사용하고 있음을 확인하였기 때문이다. 사례조사 대상은 UNESCO, Realfagstermer, PLOS, OSTI, EuroVoc이었다. 사례 조사에서 활용하고 있는 SKOS 요소를 확인하고 현재 국가과학기술표준분류의 관리현황을 검토하여 적합한 요소를 도출하여 제시하였다. 또한 ISO 11179 표준의 용어상태 관리를 참조하여 지속적인 용어관리와 서비스가 가능한 방안을 함께 제시했다.

노영희(2001)는 기계학습을 기반으로 인터넷 학술문서를 자동분류하기 위한 연구를 진행한 바 있다. 이 연구의 핵심은 kNN 분류기를 이용해서 예제기반 자동 분류기법을 적용할 경우 학습문서집단의 자질이 축소되는데 몇 퍼센트 축소함으로써 높은 성능을 얻을 수 있는지를 알아보고자 하는 것이었다. 결과적으로 kNN 분류기를 이용하여 시스템 성능을 측정한 결과 재현율과 정확율이 90% 이상 높아짐을 확인했다.

김성희, 엄재은(2008)은 기계학습을 이용한 문서 자동분류에 관한 연구를 수행한 바 있다. 이 연구에서는 수작업 분류의 한계를 극복하고, 이용자 서비스 개선을 위해 4개의 기계학습 알고리즘을 적용해서 실험하고 가장 효과적인 방법을 제안하였다. 연구대상으로는 MeSH의 8개의 주제별 범주로 각각 100개의 문헌 타이틀을 선정하였으며, 4개의 기계학습 알고리즘으로 실험을 수행하였다. 그 결과 신경망 기법과 C5.0 기법을 병행하여 사용했을 경우 단일 기법을 사용했을 경우보다 2.5%, 3.75%가 상승한 결과를 확인했다.

김해찬솔 외(2017)는 기계학습을 이용한 기록 텍스트 자동분류 사례 연구를 수행한 바 있다. 이 연구는 사례연구가 주를 이루고 있는데, 문헌 자동분류와 인공지능의 학습방식이 발전해 온 과정을 검토하였다. 또 기계학습 중 특히 지도학습 방식의 특징과 다양한 사례를 통해 기록관리 분야에 인공지능 기술을 적용해야 할 필요성에 대해 제시한 바 있다.

선행연구에서 보듯이 문헌정보학 분야에서 주제어 추천은 대부분 로그 데이터를 바탕으로 이용자 맞춤형 서비스를 제공하기 위한 용도가 주를 이룬다. 한편 국가과학기술표준분류체계를 활용한 연구는 본 연구와 마찬가지로 자동 추천을 위한 알고리즘이나 자동 분류 등에 대한 주제가

많으며, 연구에 활용된 데이터는 국가연구개발사업과 관련된 연구정보 데이터를 주로 활용하였다. 오히려 기계학습의 경우 2000년대 초반에 선행되었으나 오늘날과 같은 학습데이터 활용을 한 사례를 찾기는 어렵고, 비교적 최근의 연구에서도 사례조사가 중심이 되는 모습을 보이고 있다.

선행 연구 분석 결과 표준분류체계의 실질적인 관리 기관인 한국과학기술기획평가원의 데이터를 기반으로 한 연구는 수행된 바 없으며, 실제 대상 자료의 주제명 추천과 관련된 학습 알고리즘을 적용해 그 결과를 확인한 연구 사례도 찾기 어렵다. 따라서 본 논문은 새로운 기법을 적용한 특정 알고리즘을 개발하는 연구는 아니지만, 검증된 주제어 분류체계인 MeSH를 대상으로 관련 성과를 확인할 수 있는 주제어 추천 알고리즘을 활용했다는 점에서 의의가 있다. 또한 향후 표준분류체계의 소분류를 기술키워드화하는 과정에서 자동 추천 방식을 택할 경우, 고려할 수 있는 사항들을 도출했다는 점에서도 의의가 있다.

Ⅲ. 주제어 추천 실험

1. 활용 소프트웨어 및 하드웨어

서두에 언급한 것처럼 주제어 추천에 활용할 알고리즘은 AttentionMeSH를 그대로 활용하였으며, 관련 연구(Jin et al., 2018)에서 제시한 필수 소프트웨어도 그대로 활용하였다. 해당 소프트웨어 목록은 아래와 같다.

- Python 3.6
- torch = = 0.4.0
- numpy = = 1.13.3

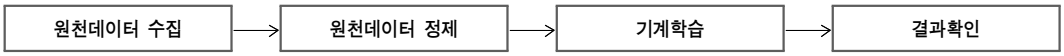
한편 원천데이터를 정제하고, 알고리즘 설치와 실험에 활용한 하드웨어 사양은 〈표 2〉와 같다.

〈표 2〉 실험용 하드웨어 사양

구분	사양
OS	Ubuntu 18.04
CPU	AMD 라이젠 7 2700X
RAM	32GB
GPU	8GB
SSD	1.5TB
HDD	2TB

2. 수행내용

주제어 추천을 위한 전체 수행 절차는 〈그림 3〉과 같다.



〈그림 3〉 주제어 추천 수행 절차

첫 번째 단계는 원천데이터 수집이다. 원천데이터는 학습 및 결과 확인을 위한 테스트 데이터세트로 활용한다. 본 연구에서 원천데이터는 한국과학기술기획평가원에서 제공한 2017년부터 2020년까지 4개년의 과제별 데이터세트를 활용하였다. 해당 원천데이터는 대외비 자료로 연도별 속성 현황을 파악하고 각각의 해당 속성에 입력된 인스턴스를 검토하여 실험에 적합한 속성과 인스턴스만을 선별하여 활용하였다. 전체 원천데이터 현황을 정리해보면 〈표 3〉과 같다.

〈표 3〉 원천데이터 현황

연번	해당연도	속성수(컬럼)	데이터수(행)
1	2017	80개	60,080건
2	2018	80개	62,842건
3	2019	83개	69,211건
4	2020	83개	72,315건
합계			264,448건

데이터세트 별로 속성수가 방대하여 여기서는 모두 기술하지 않고 〈부록 1〉에 연도별로 속성명을 비교하여 첨부하였으며, 그중 일부만 제시해 보면 〈표 4〉와 같다.

〈표 4〉 연도별 원천데이터 속성 중 일부

2017년 속성	2018년 속성	2019년 속성	2020년 속성
과제수행년도	과제수행년도	과제수행년도	과제수행년도
부처명	부처명	부처명	부처명
사업ID	사업ID	사업ID	사업ID
사업명	사업명	사업명	사업명
사업구분코드	사업구분코드	사업구분코드	사업구분코드
사업구분	사업구분	사업구분	사업구분
회계구분코드	회계구분코드	회계구분코드	회계구분코드
회계구분	회계구분	회계구분	회계구분
	녹색기술분류코드	녹색기술분류코드	녹색기술분류코드
	녹색기술분류	녹색기술분류	녹색기술분류
	27대녹색기술분류코드	27대녹색기술분류코드	27대녹색기술분류코드
	27대녹색기술분류	27대녹색기술분류	27대녹색기술분류

연도별로 속성수에 차이를 보이는 것은 해당 연도의 국가정책이나 이슈 반영 등의 영향이다. 예를 들어 2018년부터는 '녹색기술분류코드'와 같은 새로운 항목이 추가된 것을 확인할 수 있다. 연도별 속성수에 차이는 있지만 학습에 필요한 과제명, 요약문, 키워드와 같은 핵심 항목들은 공통적으로 포함되어 있어 학습에 주는 영향은 없다. 전체 데이터량은 264,778건이며, 각 데이터들은 약 80개의 속성에 분산·저장되어 있는데, 모든 속성을 기계학습에 활용하는 것은 효과가 떨어질 수 있어 본 연구에서는 연구의 핵심을 잘 표현한다고 판단되는 '과제명', '연구목표', '연구내용', '기대효과'를 학습대상으로 하였다. 이 4가지 속성은 실제 연구내용의 핵심 내용을 포괄하는 주제어가 포함될 가능성이 높다고 할 수 있다. 이 외 속성들의 경우는 〈표 4〉에서 보는 것처럼 연구내용보다는 관리에 필요한 요소들이 주를 이루어 연구내용을 직접 담고 있다고 보기 어렵다.

두 번째 단계는 원천데이터 정제로, 기계학습이 가능한 최종 데이터 형식으로 만드는 과정이다. 먼저 모든 속성에 값을 가지고 있는 데이터들이 대체로 학습 데이터로 품질이 높기 때문에 학습대상 속성인 '과제명', '연구목표', '연구내용', '기대효과' 중 'null' 값이 존재하는 데이터는 모두 삭제하였다. 그리고 이들 4개 속성과 함께 학습시킬 한글 주제어 속성에 대한 정제를 수행하였다. 여기서 한글 주제어는 학습 결과의 답이라고 할 수 있는 속성으로 학습 시 4개 속성 값에 대한 값으로 활용하였다. 주제어 속성은 같은 단어라도 띄어쓰기에 따라 다른 단어로 인지될 수 있는 가능성을 줄이기 위해 공백을 모두 제거하는 과정을 추가로 거쳤다. 또한 주제어의 경우, 4가지 핵심 속성과 마찬가지로 데이터량이 상당해 학습에 적합한 규모를 확보할 필요가 있어 본 연구에서는 100회 이상 언급된 주제어만 사용하였다. '과제명', '연구목표', '연구내용', '기대효과'의 경우는 학습시킬 질문에 해당하는 것들로 별도의 정제를 거치지 않았다. 이상의 모든 정제과정을 거쳐 최종 학습 데이터로 활용한 데이터는 4개 속성(과제명, 연구목표, 연구내용, 기대효과)에서 100회 이상 언급된 주제어를 포함하고 있는 데이터이며, 최종 정제된 학습데이터의 규모는 〈표 5〉와 같다.

〈표 5〉 최종 학습데이터의 규모

데이터 개수	한국어 주제어 개수
142,381개	1,059개

한편 〈표 5〉의 데이터는 '8:2'의 비율로 '학습데이터:결과확인데이터'로 구분하여 활용했다. 즉 142,381개의 데이터 중 113,905건의 데이터를 학습데이터로, 나머지 28,476 데이터는 학습을 마친 후 결과 확인을 위한 데이터로 활용했다.

세 번째 단계는 기계학습 과정으로 정제가 완료된 데이터 중 학습용 데이터세트를 대상으로 AttentionMeSH 알고리즘을 적용하였다. 마지막 단계는 결과확인으로 AttentionMeSH의 결과 확인 절차에 따라 Micro-F, MiP, MiR 값을 확인했다.

3. 실험결과

일반적으로 임계치(threshold)가 0.5 이상이면 정답을 제시할 확률이 가장 높은 신뢰할 수 있는 결과로 볼 수 있다. 실험 결과를 정리한 〈표 6〉에서 제시된 바와 같이, 이번 실험의 결과도 임계값이 0.5일 때 MiF가 가장 높은 것을 확인할 수 있었다. MiF 0.6377이라는 결과는 관련 연구에서 제시된 AttentionMeSH 알고리즘 완성 후 확인한 결과값 0.684에는 다소 미치지 못하는 수치지만, MeSH가 아닌 다른 원천데이터를 활용하고 얻은 결과라는 점을 감안했을 때 결코 낮은 수치로 보기는 어렵다.

〈표 6〉 자동분류 결과

threshold	MiF	MiP	MiR
0.1	0.5136	0.3920	0.7449
0.2	0.5842	0.5012	0.7000
0.3	0.6146	0.5723	0.6636
0.4	0.6311	0.6295	0.6327
0.5	**0.6377**	**0.6783**	**0.6017**
0.6	0.6354	0.7200	0.5686
0.7	0.6265	0.7617	0.5321
0.8	0.6061	0.8077	0.4850
0.9	0.5548	0.8556	0.4105

〈그림 4〉, 〈그림 5〉는 학습을 종료한 알고리즘을 활용하여 실험 데이터세트를 검증한 결과의 일부이다. 이는 임계값을 0.5로 주고 시험 데이터로 논문 제목과 초록을 입력했을 경우 자동으로 추천받은 주제어와 MiF 값을 보여준다.

```
title : 머신러닝 기법을 이용한 대졸자 취업예측 모형
text : 본 연구는 머신러닝의 랜덤 포리스트 기법을 이용하여 대졸자의 취업 여부와 취업의 질을 예측하는 모형을 제시한다. 전통적 회귀분석에서는 설명변수의 외생성이나 오차항 분포에 대한 제약이 있지만 머신러닝 접근법은 이러한 제약에서 상대적으로 자유로운 편이다. 본 연구에서 사용한 예측인자에는 대졸자의 객관적 특성 변수뿐 아니라 취업관련 프로그램 참여 여부, 일자리 선택 시 고려사항, 감정빈도 변수 등 응답자가 주관적으로 평가하는 특성까지 포함되어 있다. 분석결과를 보면, 객관적 및 주관적 예측인자를 모두 사용하는 모형이 객관적 예측인자만 사용한 모형에 비해 취업여부와 취업의 질을 예측하는 모형 모두에서 예측성과가 더 우수한 것으로 나타났다. 랜덤 포리스트 기법을 사용하여 예측인자의 상대적 중요성을 파악한 결과, 취업여부 모형에서는 가구주 여부, 부모동거 여부와 같은 객관적 변수뿐 아니라 주관적 변수인 감정빈도 변수도 중요한 영향을 미쳤다. 취업의 질에 있어서도 학교유형이나 전공계열과 같은 객관적 변수는 물론 주관적 감정빈도 변수가 중요한 영향을 미치는 것으로 나타났다.
result : ['예측모형', '머신러닝']
score : [0.998863697052002, 0.9419066905975342]
```

〈그림 4〉 테스트 데이터로 점검한 주제어 추천 결과 1

```
title : 담배 흡연이 생쥐의 흉선 및 비장의 면역 기능에 미치는 영향
text : 본 연구에서는 담배필터를 통해 호흡기로 노출되는 주류연 흡연 장치를 구현하여 직접 흡연에 의한 주류연의 노출이 면역시스템에 미치는 영향을 생쥐모델을 이용하여 규명하고자 하였다. 실험동물(n = 15)은 대조군(sham군)과 담배 2개비/일 노출군 및 4개비/일 노출군으로 각각 나누고, 7일 동안 주류연을 노출시킨 후 체중변화와 면역기관 지수를 측정하였다. 주류연 노출이 면역세포의 기능에 미치는 영향을 확인하기 위해, lipopolysaccharide(LPS) 또는 concanavalinA(ConA) 자극에 대한 비장세포와 흉선세포의 증식력과 nitric oxide(NO) 분비량을 측정하였다. 실험결과, 4개비/일 고용량 노출군의 흉선, 비장, 간 무게가 대조군에 비해 각각 0.3, 0.2 및 0.2배 유의적으로 감소하였고(P<0.05), 흉선세포 생존율과 B세포 및 T세포 증식력이 감소하였다. 또한, LPS에 의한 비장세포의 NO 분비능도 감소하였다. 본 연구는 주류연 노출이 흉선과 비장에 직접적인 손상을 주며, 자극원에 대한 면역세포의 증식력과 NO 분비 등 방어능력을 유의적으로 약화시킨다는 것을 보여주었다. 이는 흡연과 관련된 면역질환 발생이 주류연의 노출에 대한 직접적인 면역기관의 손상과 자극에 대한 방어 기능의 약화가 관련되어 있음을 제시한다.
result : ['동물모델']
score : [0.7734372615814209]
```

〈그림 5〉 테스트 데이터로 점검한 주제어 추천 결과 2

〈그림 4〉에서 도출된 추천 키워드는 '예측모형'과 '머신러닝'이다. 해당 연구가 정식논문으로 출판된 결과를 찾아보면 저자가 작성한 키워드는 '취업예측', '취업의 질', '머신러닝', '랜덤 포리스트'로 총 4가지이다. 〈그림 5〉에서 도출한 키워드는 '동물모델'이며, 해당 정식논문에서 제시한 키워드는 국문없이 영문으로 'Smoke', 'Thymocyte', 'Splenocyte', 'Cell Viability', 'Mainstream Smoke'이다. 결론적으로 〈그림 4〉의 논문은 하나의 키워드만 일치하고, 〈그림 5〉의 논문에서는 일치하는 키워드가 없음을 확인했다. 영어로 표현된 키워드를 국문으로 변경하더라도 일치하는 키워드는 존재하지 않는다. 이 결과는 본 연구에서 수행한 실험의 기준이 현재의 소분류 용어를 기준으로 하고 있기 때문이다. 즉, 소분류에 존재하지 않는 용어는 추출이 어렵고, 현재의 소분류체계가 연구자와 연구 자체의 특징을 다양하게 반영하고 있지 못하다는 반증이기도 하다.

Ⅳ. 결론 및 제언

본 연구의 목적은 국가과학기술표준분류체계의 소분류를 기술키워드화하고 해당 키워드를 도출하는 방법 중 하나로 기존에 타 연구에서 검증된 알고리즘에 기반한 자동 추천 방식을 적용해보고 그 결과를 분석해보는 것이다. 이는 표준분류체계 관리 운영 주체인 한국과학기술기획평가원이 도입할 수 있는 방법 중 하나로 분류체계 개편 초기에 기술키워드의 양적인 충당 문제와 개별 연구자들이 표준용어를 선택하는 데 겪는 어려움을 해결할 수 있도록 하기 위함이다. 이상의 연구 목적을 달성하기 위해, 본 연구에서는 BioASQ 대회를 통해 그 성능이 확인된 AttentionMeSH 알고리즘에 한국과학기술기획평가원 연구현황 파일의 원천데이터를 정제·적용해 기계학습을 수행하고, 임계값 0.5 수준에서 자동 추천되는 주제어가 대상 자료를 표현하는데 적합한 키워드로 제시됨을 확인했다.

AttentionMeSH 알고리즘의 주제어 추천 적합성 결과 외에도 이번 실험을 진행하는 과정에서

향후 기술키워드 확보를 위한 자동 추천 알고리즘 적용의 신뢰성을 높이기 위해서는 기계학습에 활용할 충분한 수준의 데이터세트 확보가 필수적이라는 것을 확인할 수 있었다. 결과로 도출된 추천 키워드들은 대체로 대상 연구 자료를 대표할 수 있는 용어들이 많이 제시되었지만, 동일하거나 유사한 용어가 반복적으로 제안된 경우가 많았다. 이러한 결과가 나온 주요 원인을 살펴보면, 실제로 학습 데이터세트에 사용된 용어의 개수(62,236개, 중복을 제거한 학습용 데이터세트의 수)가 전체 데이터개수(142,381개)에 비해 적었다는 점이다. 즉 국문키워드에 사용된 용어들이 동일 용어를 반복적으로 사용하는 경우가 많아 실제 학습 데이터로 활용할 수 있는 용어의 양이 극히 적었다는 문제가 있었다. 실제로 〈표 7〉은 이번 실험에서 정제한 한글 키워드 중 자주 등장하는 용어를 정리한 것이다.

〈표 7〉 자주 등장하는 키워드 빈도수

순위	해당 키워드	빈도수
1	인공지능	5,436
2	빅데이터	4,131
3	딥러닝	3,680
4	사물인터넷	3,247
5	바이오마커	2,097
6	기계학습	2,060
7	기후변화	1,524
8	머신러닝	1,470
9	가상현실	1,447

〈표 7〉에 등장하는 용어들은 원천데이터에서 학습에 적합한 데이터들로 정제한 후라는 점은 감안할 필요가 있다. 그럼에도 동일 용어 사용이 계속 반복되고, 다양성이 부족한 결과가 나온 것은 기계학습과 자동추천에 걸림돌로 작용한다. 물론 〈표 7〉에 등장하는 용어들은 학습에 사용된 원천데이터가 비교적 최근의 자료들이고 다양한 학문 분야에서 최근에 공통적으로 자주 사용되며 화두가 되고 있는 분야의 용어라는 점을 고려할 필요가 있다. 또한 기계학습이라는 목적을 떠나서 현재 표준분류체계의 소분류 체계가 갖는 문제점도 감안할 필요가 있다. 즉 확정된 소분류 용어를 사용해야 하므로 분류 대상 연구가 포함하는 다양한 주제를 연구자가 표현하기 어려운 구조라는 점이다. 한국과학기술기획평가원에서 기존의 소분류 체계를 기술키워드화 하고자 하는 목적도 이런 문제점을 극복하기 위한 노력으로 볼 수 있다.

본 연구 결과에 중요한 영향을 준 또 다른 요인 중 하나는 한국과학기술기획평가원에서 관리하는 속성이 약 80개 정도로 방대함에도 불구하고, 실제로 연구내용을 제대로 표현할 수 있는 속성의 수는 제한적이고 각 속성에 내용이 입력되지 않고 비어 있는 값들이 상당수 존재한다는 점이다. 따라서 향후 기계학습을 통한 자동추천을 수행한다고 할 경우 현재보다 다양한 속성들을 모두 학습범위

로 포함시키는 것이 바람직하다. 또한 기술키워드 체계로 변환되고 나서 다양한 분야의 이용자들을 통해 수집된 기술 용어들이 확보되면 이를 계속 학습 데이터로 활용하는 것을 추천한다. 이외에도 학습의 품질을 보장하기 위해서는 주제어 언급 횟수가 많은 용어들은 조정하여 편향을 방지하는 작업이 필요하며, 표준용어 외에도 지속적인 용어 관리를 통해 유의어, 대체어 등 다양한 관련어 데이터를 추가하는 것도 필요하다. 정리하면 현재 표준분류체계 용어를 학습데이터로 활용하는데는 두 가지 한계점이 존재한다. 하나는 학습데이터의 양적부족이다. 물론 기존 소분류체계에서 활용하는 용어는 극히 제한적일 수 밖에 없지만 그럼에도 동일용어가 모든 학문분야에서 공통, 반복적으로 등장하는 것은 학습데이터의 품질을 떨어뜨린다고 볼 수 있다. 향후 학습을 위해서는 기존 소분류 용어 외에 제목, 초록 등 논문의 핵심적인 내용을 담고 있는 속성에서 주제어를 도출하여 활용하는 방안과 국내외 다양한 과학기술용어, 표준분류체계의 용어를 함께 활용하는 것이 필요하다. 두 번째는 데이터 자체의 품질문제로, 대표적인 사례로 다수의 비어있는 값들을 들 수 있다. 향후 기계학습을 위한 데이터 입수와 활용을 위해서는 원천데이터인 학습데이터의 품질관리를 위한 방안이 필요할 것으로 보인다.

본 연구가 갖는 가장 큰 한계는 기계학습 결과로 추천된 주제어가 결국 학습데이터의 범주를 벗어나지 못한다는 점이다. 결론적으로 만약 연구자가 이 시스템을 활용하거나 관리 주체인 한국과학기술기획평가원이 활용한다고 하더라도 주제어 추천은 학습 데이터세트의 범주를 벗어나지 못한다는 것을 의미한다. 그럼에도 본 연구가 갖는 의의는 향후 위에 언급한 개선점들을 참조하고 보다 다양한 알고리즘 혹은 한국과학기술기획평가원의 용도에 맞는 새로운 알고리즘을 직접 개발하고 보완하는 방식을 거치면 사람에 의한 추천 외에 보조적인 수단으로 활용 가능함을 확인했다는 점에 있다고 볼 수 있다. 기계학습에 의한 주제어 추천은 현재 본 사업으로 진행되고 있지 않으며 향후 선택할 수 있는 방안 중 하나라는 점을 다시 한 번 지적하며, 초기에는 사람의 개입과 검증, 양적 충당을 위한 노력이 선행되어야 한다.

참 고 문 헌

과학기술정보통신부 (2019) 국가과학기술표준분류체계 개정타당성 평가대상 선정결과(안) 및 중장기 개선방향. 과학기술정보통신부 성과평가정책국 과학기술정보과.

김광영, 곽승진 (2010). 개인화 검색시스템에 관한 연구 - 과학기술학회마을을 중심으로 -. 한국도서관·정보학회지, 41(1), 149-165. https://doi.org/10.16981/kliss.41.1.201003.149

김광영, 곽승진 (2011). 주제분류 기반의 개인화 검색시스템에 관한 연구. 한국문헌정보학회지, 45(4), 77-102. https://doi.org/10.4275/KSLIS.2011.45.4.077

김성희, 엄재은 (2008). 기계학습을 이용한 문서 자동분류에 관한 연구. 정보관리연구, 39(4), 47-66.

김윤정, 신동구, 정회경 (2021). 머신러닝을 이용한 R&D과제의 연구분야 추천 서비스. 한국정보통신학회논문지, 25(12), 1809-1816. https://doi.org/10.6109/jkiice.2021.25.12.1809

김해찬솔, 안대진, 임진희, 이해영 (2017). 기계학습을 이용한 기록 텍스트 자동분류 사례 연구. 정보관리학회지, 34(4), 321-344. https://doi.org/10.3743/KOSIM.2017.34.4.321

노영희 (2001). 기계학습을 기반으로 한 인터넷 학술문서의 효과적 자동분류에 관한 연구. 한국도서관·정보학회지, 32(3), 307-330.

송민선, 박진호 (2021). 국가과학기술표준분류체계 용어 관리를 위한 SKOS 기반 메타데이터 요소 개발 연구. 한국비블리아학회지, 32(4), 67-88. https://doi.org/10.14699/kbiblia.2021.32.4.067

이소영, 정영미 (2006). 웹 포털 이용자 로그 데이터에 기반한 개인화 검색 서비스 모형의 설계 및 평가. 정보관리학회지, 23(4), 179-196. http://doi.org/10.3743/KOSIM.2006.23.4.179

조현양 (2017). 자동분류기반 성격 유형별 도서추천시스템 개발을 위한 실험적 연구. 한국도서관·정보학회지, 48(2), 215-236. https://doi.org/10.16981/kliss.48.201706.215

조현양 (2020). 대체자료 선정을 위한 이용자 참여형 도서 추천 큐레이션 플랫폼 설계. 한국문헌정보학회지, 54(3), 41-69. https://doi.org/10.4275/KSLIS.2020.54.3.041

최종윤, 한혁, 정유철 (2020). 국가 과학기술 표준분류 체계 기반 연구보고서 문서의 자동 분류 연구. 한국산학기술학회논문지, 21(1), 169-177. http://10.5762/KAIS.2020.21.1.169

한국과학기술기획평가원 (2019) 과학기술기획 및 혁신정책 활용도 제고를 위한 KISTEP 미래예측 역할 재정립 연구. 충청북도: 한국과학기술기획평가원.

한희준, 최윤수, 최성필 (2018). 개인 관심분야 추적기법을 이용한 과학기술정보 개인화에 관한 연구. 한국문헌정보학회지, 52(3), 5-33. http://10.4275/KSLIS.2018.52.3.005

BioAsQ [n.d.]. Challenges - Tasks 6a, 6b - Year 6. Available: http://bioasq.org/participate/challenges_year_6

BioASQ (n.d.). Sixth Challenge Winners. Available: http://bioasq.org/participate/sixth-challenge-winners

Jin, Q., Dhingra, B., Cohen, W., & Lu, X. (2018). Attentionmesh: simple, effective and interpretable automatic mesh indexer. Proceedings of the 6th BioASQ Workshop a Challenge on Large-scale Biomedical Semantic Indexing and Question Answering, 47-59. https://doi.org/10.18653/v1/W18-5306

• **국한문 참고문헌의 영문 표기**
(English translation / Romanization of references originally written in Korean)

Cho, Hyun Yang (2017). A experimental study on the development of a book recommendation system using automatic classification, based on the personality type. Journal of Korean Library and Information Science Society, 48(2), 215-236. https://doi.org/10.16981/kliss.48.201706.215

Cho, Hyun Yang (2020). Design of the curation platform for user-participated book recommendation system of selecting on alternative material for the disabled. Journal of the Korean Society for Library and Information Science, 54(3), 41-69. https://doi.org/10.4275/KSLIS.2020.54.3.041

Choi, Jong-Yun, Hahn, Hyuk, & Jung, Yu Chul (2020). Research on text classification of research reports using Korea national science and technology standards classification codes. Journal of the Korea Academia-Industrial Cooperation Society, 21(1), 169-177. http://10.5762/KAIS.2020.21.1.169

Han, Hee-Jun, Choi, Yunsoo, & Choi, Sung-Pil (2018). A study on personalization of science and technology information by user interest tracking technique. Journal of the Korean Society for Library and Information Science, 52(3), 5-33. http://10.4275/KSLIS.2018.52.3.005

Kim, Hae Chan Sol, Ahn, Dae Jin, Yim, Jin Hee, & Rieh, Hae-Young (2017). A study on automatic classification of record text using machine learning. Journal of the Korean Society for Information Management, 34(4), 321-344. https://doi.org/10.3743/KOSIM.2017.34.4.321

Kim, Kwang-Young & Kwak, Seung-Jin (2010). A study of personalized retrieval system through society of Korean journal articles of science and technology. Journal of Korean Library and Information Science Society, 41(1), 149-165. http://10.16981/kliss.41.1.201003.149

Kim, Kwang-Young & Kwak, Seung-Jin (2011). A study on personalized search system based on subject classification. Journal of the Korean Society for Library and Information Science, 45(4), 77-102. http://dx.doi.org/10.4275/KSLIS.2011.45.4.077

Kim, Sunghee & Eom, Jae-Eun (2008). A study on the documents's automatic classification using machine learning. Journal of Information Management, 39(4), 47-66.

Kim, Yunjeong, Shin, Donggu, & Jung, Hoikyung (2021). Recommendation system for research field of R&D project using machine learning. Journal of the Korea Institute of Information and Communication Engineering, 25(12), 1809-1816. http://doi.org/10.6109/jkiice.2021.25.12.1809

Korea Institute of S&T Evaluation and Planning (2019). A Study on Reestablishing the Role of KISTEP for Predicting the Future to Enhance the Utilization of Science and Technology Planning and Innovation Policy. Chung-cheong bukdo: Korea Institute of S&T Evaluation and Planning.

Lee, Soyoung & Chung, Young-Mee (2006). Design and evaluation of a personalized search service model based on web portal user activities. Journal of the Korean Society for Information Management, 23(4), 179-196. http://doi.org/10.3743/KOSIM.2006.23.4.179

Ministry of Science and Technology Information and Communication (2019). Selection Results (Draft) for the Revision Feasibility Evaluation of the National Science and Technology Standards Classification System and the Long-term Improvement Direction. Ministry of Science and ICT, Performance Evaluation Policy Bureau, Science and Technology Information Division.

Noh, Young-Hee (2001). The study on the effective automatic classification of internet document using the machine learning. Journal of Korean Library and Information Science Society, 32(3), 307-330.

Song, Min Sun & Park, Jin Ho (2021). A study on development of SKOS-based metadata elements for managing keywords in the national science and technology standard classification system. Journal of the Korean Biblia Society for Library and Information Science, 32(4), 67-88. https://doi.org/10.14699/kbiblia.2021.32.4.067

[부록 1] KISTEP 제공 원천데이터의 연도별 속성수 비교

	2017	2018	2019	2020
1	과제수행년도	과제수행년도	과제수행년도	과제수행년도
2	부처명	부처명	부처명	부처명
3	사업ID	사업ID	사업ID	사업ID
4	사업명	사업명	사업명	사업명
5	사업구분코드	사업구분코드	사업구분코드	사업구분코드
6	사업구분	사업구분	사업구분	사업구분
7	회계구분코드	회계구분코드	회계구분코드	회계구분코드
8	회계구분	회계구분	회계구분	회계구분
9	과제고유번호	과제고유번호	과제고유번호	과제고유번호
10	(기관)세부과제번호	신규계속구분	신규계속구분코드	신규계속구분코드
11	신규계속구분코드	내역사업명	신규계속구분	신규계속구분
12	신규계속구분	과제명-국문	내역사업명	내역사업명
13	내역사업명	총연구기간-시작년월일	대과제명	대과제명
14	대과제명	총연구기간-종료년월일	내역사업명(구.대과제명)	내역사업명(구.대과제명)
15	내역사업명(구.대과제명)	당해연구기간-시작년월일	과제명-국문	과제명-국문
16	과제명-국문	당해연구기간-종료년월일	총연구기간-시작년월일	총연구기간-시작년월일
17	총연구기간-시작년월일	녹색기술분류코드	총연구기간-종료년월일	총연구기간-종료년월일
18	총연구기간-종료년월일	녹색기술분류	당해연구기간-시작년월일	당해연구기간-시작년월일
19	당해연구기간-시작년월일	27대녹색기술분류코드	당해연구기간-종료년월일	당해연구기간-종료년월일
20	당해연구기간-종료년월일	27대녹색기술분류	녹색기술분류코드	녹색기술분류코드
21	주관/협동구분	연구개발단계코드(변경)	녹색기술분류	녹색기술분류
22	연구개발단계코드	연구개발단계(변경)	27대녹색기술분류코드	27대녹색기술분류코드
23	연구개발단계	연구수행주체코드	27대녹색기술분류	27대녹색기술분류
24	연구수행주체코드	연구수행주체	연구개발단계코드	연구개발단계코드
25	연구수행주체	과학기술표준분류코드1-대	연구개발단계	연구개발단계
26	과학기술표준분류코드1-대	과학기술표준분류1-대	연구수행주체코드	연구수행주체코드
27	과학기술표준분류1-대	과학기술표준분류코드1-중	연구수행주체	연구수행주체
28	과학기술표준분류코드1-중	과학기술표준분류1-중	과학기술표준분류코드1-대	과학기술표준분류코드1-대
29	과학기술표준분류1-중	과학기술표준분류코드1-소	과학기술표준분류1-대	과학기술표준분류1-대
30	과학기술표준분류코드1-소	과학기술표준분류1-소	과학기술표준분류코드1-중	과학기술표준분류코드1-중
31	과학기술표준분류1-소	과학기술표준분류가중치1	과학기술표준분류1-중	과학기술표준분류1-중
32	과학기술표준분류가중치1	과학기술표준분류코드2-대	과학기술표준분류코드1-소	과학기술표준분류코드1-소
33	과학기술표준분류코드2-대	과학기술표준분류2-대	과학기술표준분류1-소	과학기술표준분류1-소
34	과학기술표준분류2-대	과학기술표준분류코2-중	과학기술표준분류가중치1	과학기술표준분류가중치1
35	과학기술표준분류코2-중	과학기드술표준분류2-중	과학기술표준분류코드2-대	과학기술표준분류코드2-대
36	과학기드술표준분류2-중	과학기술표준분류코드2-소	과학기술표준분류2-대	과학기술표준분류2-대
37	과학기술표준분류코드2-소	과학기술표준분류2-소	과학기술표준분류코2-중	과학기술표준분류코2-중
38	과학기술표준분류2-소	과학기술표준분류가중치2	과학기드술표준분류2-중	과학기드술표준분류2-중
39	과학기술표준분류가중치2	과학기술표준분류코드3-대	과학기술표준분류코드2-소	과학기술표준분류코드2-소
40	과학기술표준분류코드3-대	과학기술표준분류3-대	과학기술표준분류2-소	과학기술표준분류2-소
41	과학기술표준분류3-대	과학기술표준분류코드3-중	과학기술표준분류가중치2	과학기술표준분류가중치2

	2017	2018	2019	2020
42	과학기술표준분류코드3-중	과학기술표준분류3-중	과학기술표준분류코드3-대	과학기술표준분류코드3-대
43	과학기술표준분류3-중	과학기술표준분류코드3-소	과학기술표준분류3-대	과학기술표준분류3-대
44	과학기술표준분류코드3-소	과학기술표준분류3-소	과학기술표준분류코드3-중	과학기술표준분류코드3-중
45	과학기술표준분류3-소	과학기술표준분류가중치3	과학기술표준분류3-중	과학기술표준분류3-중
46	과학기술표준분류가중치3	적용분야코드1	과학기술표준분류코드3-소	과학기술표준분류코드3-소
47	적용분야코드1	적용분야1	과학기술표준분류3-소	과학기술표준분류3-소
48	적용분야1	적용분야가중치1	과학기술표준분류가중치3	과학기술표준분류가중치3
49	적용분야가중치1	적용분야코드2	적용분야코드1	적용분야코드1
50	적용분야코드2	적용분야2	적용분야1	적용분야1
51	적용분야2	적용분야가중치2	적용분야가중치1	적용분야가중치1
52	적용분야가중치2	적용분야코드3	적용분야코드2	적용분야코드2
53	적용분야코드3	적용분야3	적용분야2	적용분야2
54	적용분야3	적용분야가중치3	적용분야가중치2	적용분야가중치2
55	적용분야가중치3	6T관련기술코드-대	적용분야코드3	적용분야코드3
56	6T관련기술코드-대	6T관련기술-대	적용분야3	적용분야3
57	6T관련기술-대	6T관련기술코드-중	적용분야가중치3	적용분야가중치3
58	6T관련기술코드-중	6T관련기술-중	6T관련기술코드-대	6T관련기술코드-대
59	6T관련기술-중	6T관련기술코드-소	6T관련기술-대	6T관련기술-대
60	6T관련기술코드-소	6T관련기술-소	6T관련기술코드-중	6T관련기술코드-중
61	6T관련기술-소	중점과학기술코드-대	6T관련기술-중	6T관련기술-중
62	기술수명주기코드	중점과학기술-대	6T관련기술코드-소	6T관련기술코드-소
63	기술수명주기	중점과학기술코드-중	6T관련기술-소	6T관련기술-소
64	세부과제성격코드	중점과학기술-중	중점과학기술코드-대	중점과학기술코드-대
65	세부과제성격	중점과학기술코드-소	중점과학기술-대	중점과학기술-대
66	정부연구비(원)	중점과학기술-소	중점과학기술코드-중	중점과학기술코드-중
67	경제사회목적코드	기술수명주기코드	중점과학기술-중	중점과학기술-중
68	경제사회목적	기술수명주기	중점과학기술코드-소	중점과학기술코드-소
69	국가전략기술코드-대	세부과제성격코드	중점과학기술-소	중점과학기술-소
70	국가전략기술-대	세부과제성격	기술수명주기코드	기술수명주기코드
71	국가전략기술코드-중	정부연구비(원)	기술수명주기	기술수명주기
72	국가전략기술-중	경제사회목적코드	세부과제성격코드	세부과제성격코드
73	국가전략기술코드-소	경제사회목적	세부과제성격	세부과제성격
74	국가전략기술-소	공동연구협력유형구분코드	정부연구비(원)	정부연구비(원)
75	공동연구협력유형구분코드	공동연구협력유형구분	경제사회목적코드	경제사회목적코드
76	요약문_연구목표	요약문_연구목표	경제사회목적	경제사회목적
77	요약문_연구내용	요약문_연구내용	공동연구협력유형구분코드	공동연구협력유형구분코드
78	요약문_기대효과	요약문_기대효과	공동연구협력유형구분	공동연구협력유형구분
79	요약문_한글키워드	요약문_한글키워드	요약문_연구목표	요약문_연구목표
80	요약문_영문키워드	요약문_영문키워드	요약문_연구내용	요약문_연구내용
81			요약문_기대효과	요약문_기대효과
82			요약문_한글키워드	요약문_한글키워드
83			요약문_영문키워드	요약문_영문키워드