ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
https://www.jiisonline.org
J Intell Inform Syst 2023 September: 29(3): 37~55 http://dx.doi.org/10.13088/jiis.2023.29.3.037

Sentence BERT를 이용한 내용 기반 국문 저널추천 시스템

김용우
한양대학교 기술경영전문대학원 기술경영학과
(ywkim@hanyang.ac.kr)

김대영
디플래닉스(주)
(daeyoung.kim@dplanex.com)

서현희
디플래닉스(주)
(hyunhee.seo@dplanex.com)

김영민
한양대학교 기술경영전문대학원 기술경영학과
(yngmnkim@hanyang.ac.kr)

전자저널의 발전과 다양한 융복합 연구들이 생겨나면서 연구를 게시할 저널의 선택은 신진 연구자들은 물론 기존 연구자들에게도 새로운 문제로 떠올랐다. 논문의 수준이 높더라도 논문의 주제와 저널 범위의 불일치로 인해 게재가 거부될 수 있기 때문이다. 이러한 문제를 해결하기 위해 연구자의 저널 선정을 돕기 위한 연구는 영문 저널을 대상으로는 활발하게 이루어졌으나 한국어 저널을 대상으로 한 연구는 그렇지 못한 실정이다. 본 연구에서는 한국어 저널을 대상으로 투고할 저널을 추천하는 시스템을 제시한다. 첫 번째 단계는 과거 저널에 게재된 논문들의 초록을 SBERT (Sentence-BERT)를 이용하여 문서 단위로 임베딩하고 새로운 문서와 기존 게재논문의 유사도를 비교하여 저널을 추천하는 것이다. 다음으로 초록의 유사도 여부, 키워드 일치 여부, 제목 유사성을 고려하여 추천할 저널의 순서가 결정되고, 저널별로 구축된 단어 사전을 이용하여 선순위 추천 저널과 유사한 저널을 찾아 추천 리스트에 추가하여 추천 다양성을 높인다. 이러한 방식으로 구축된 추천 시스템을 평가한 결과 Top-10 정확도 76.6% 수준으로 평가되었으며, 추천 결과에 대한 사용자의 평가를 요청하고 추천 결과의 유효성을 확인하였다. 또한, 제안된 프레임워크의 각 단계가 추천 정확도를 높이는 데에 도움이 된다는 결과를 확인하였다. 본 연구는 그동안 활발히 이루어지지 않았던 국문 학술지 추천에 대한 새로운 접근을 제시한다는 점에서 학술적 의의가 있으며, 제안된 기능을 문서와 저널 보유상태에 따라 변경하여 손쉽게 서비스에 적용할 수 있다는 점에서 실무적인 의의를 가진다.

주제어 : 딥러닝, 문서유사도, 추천시스템, 논문, SBERT

논문접수일 : 2023년 6월 14일 논문수정일 : 2023년 7월 6일 게재확정일 : 2023년 7월 17일
원고유형 : Regular Track 교신저자 : 김영민

1. 개요

IT 기술의 급속한 발달과 전자저널화는 많은 양의 학술논문이 쉽게 출판되고 공유될 수 있도록 했다. 덕분에 생산자의 출판 과정이 효율화됨은 물론 독자들도 시간과 공간의 제약 없이 편리하게 학술논문에 접근할 수 있게 되었다 (신은자, 2001). 나아가 전자저널 플랫폼이 등장하면서 여러 저널의 데이터가 중앙집중형으로 수집·저장되었으며, 덕분에 학술논문을 이용하여 새로운 가치를 창출하기 위한 통합적 분석이 가능해졌다.

한편, 융복합 연구가 점점 활발해지고 논문을 출판하는 저널 역시 다양해지면서, 자신의 연구를 적합한 저널에 출판하는 과정은 점점 더 어려워지고 있다. 특히 제출 경험이 거의 없는 신진 학자들의 경우 적합한 저널이나 학회를 선택하는 것이 상대적으로 어렵다. 또한, 기존 연구자들 역시 연구주제의 변경에 따라서 기존에는 알지 못했던 저널이나 학회를 찾아 투고해야 하는 경우가 있다. 이러한 상황에서 연구자가 투고할만한

저널을 추천할 수 있는 저널추천 시스템이 도움이 될 수 있다. 이러한 기능을 할 수 있는 추천 시스템에 관한 연구는 기존에도 많이 이루어졌다. 그러나 DBLP[1], AMiner[2], APS[3]와 같이 연구에 필요한 공개된 학술 데이터 세트가 풍부한 영문 저널을 대상으로 이루어져 한국어를 사용하는 저널을 대상으로 한 연구는 미비한 실정이다. 국내의 경우에도 2,700종 이상의 KCI 등재(후보)지가 존재하여 연구자가 자신이 투고할 저널을 결정하기 위해서 인접 연구 분야에 속하는 모든 저널의 정보를 확인하기는 어렵다. 본 연구에서는 한국과학기술정보연구원(KISTI, Korea Institute of Science and Technology Information)에서 제공하는 한국어 논문 데이터셋을 이용하여 추천 시스템을 구축하고 그 결과를 평가하였다. 추천 시스템을 구축하기 위해 문서의 유사도를 계산하는 방법을 이용하며, 유사도 계산을 위해 한국어 데이터셋으로 훈련된 Sentence-BERT 모델을 이용하여 기존 출판된 논문의 초록과 사용자가 입력한 초록의 임베딩(embedding)을 구하고 코사인 유사도를 구하는 방식을 이용하였다. 초록 외에도 논문의 제목, 논문의 키워드 등을 입력하면 더욱 적합한 저널을 추천할 수 있도록 하였다.

제2장에서는 Yang & Davison (2012)의 분류에 따라 관련된 연구와 저널추천을 위해 사용된 방법에 관해서 서술한다. 제3장에서는 사용된 데이터셋에 대해 살펴본다. 제4장에서는 제안하는 추천 시스템 프레임워크와 사용된 방법론에 대해 살펴보고, 제5장에서는 시스템의 평가 방식과 평가결과에 대해 서술한다. 마지막으로 제6장에서는 결과와 한계점에 대해 논의한다.

1) https://dblp.org
2) http://www.arnetminer.org/data
3) https://journals.aps.org/datasets

2. 관련 연구

2.1 협업 필터링 기반 추천

협업 필터링은 추천 시스템에서 가장 많이 사용되는 기술 중 하나로, 사용자 혹은 아이템 간 유사성을 기반으로 추천하는 방식이다(최슬비 등, 2016). 추천 대상이 되는 목표 사용자와 유사한 선호도를 보이는 다른 사용자를 찾아서 목표 사용자가 선호할만한 아이템을 찾아 추천하는 사용자 기반 협업 필터링, 기 선호를 보이는 아이템과 유사도를 보이는 다른 아이템을 추천 대상으로 선정하는 아이템 기반 협업 필터링이 있다(최인복 & 이재동, 2009). 협업 필터링은 제품, 음악, 영화 등 다양한 종류의 추천 시스템에서 사용되고 있다(손지은 등, 2015). 마찬가지로 저널추천 시스템에서도 협업 필터링 기법이 이용되었다. 예컨대 Alhoori & Furuta (2017)은 연구자의 논문 열람 행태를 바탕으로 추천 시스템을 고안하였다. 이외에도 여러 연구자들이 이 방법을 이용하여 연구자들이 직면한 다양한 문제를 해결하기 위해 노력하였다(Yang et al., 2012; Liang et al., 2016; 손연빈 등, 2019).

2.2 네트워크 기반 추천

이 방법은 그래프 이론을 기반으로 하는 사용자 네트워크를 생성하고 추천에 활용하는 방식이다. 협업 필터링 기반의 추천방식에서는 신규 사용자나 아이템이 생성되는 경우 다른 사용자나 아이템과의 유사성을 계산하기 어렵다. 이러한 문제를 해결하기 위해 사용자 간 네트워크를

생성하여 유사도를 간접적으로 계산한다. 저널 추천을 위해서는 일반적으로 공동 연구자, 저자, 인용 네트워크 등을 그래프 형태로 구성하여 모델을 구성한다. 예를 들어 Luong et al. (2012a)는 논문의 저자 및 공동 저자와 비슷한 영역에서 연구하는 연구자들의 정보를 고려하는 소셜 네트워크 기반 저널추천을 제안하였다. 이후에도 유사한 논문 제출 패턴을 보인 연구자와의 관계를 이용하거나(Luong et al., 2012b), 공동출판 빈도 및 학업 수준이 유사한 다른 연구자들과의 관계를 이용하여 추천하는 연구들이 이루어졌다(Chen et al., 2015; Yu et al., 2018).

2.3 내용 기반 추천

내용 기반으로 추천하는 방식은 논문의 내용에 중점을 두고 저널을 추천한다. 구체적으로는 제출하고자 하는 논문의 초록과 저널에 게재된 각 논문이 얼마나 유사한지를 판단하는 것이다. 예를 들어 Kang et al. (2015)는 논문 제목, 초록, 키워드, 연구 분야를 이용하여 Elsevier 출판사의 저널을 제안하는 Elsevier Journal Finder[4]라는 저널 추천 시스템을 고안하였다. Feng et al. (2019)는 PubMed의 저널 목록을 추천하는 Pubmender라는 추천 시스템을 제안하였다. 이외에도 Springer Journal Suggester[5], Wiley Journal Finder[6], Taylor & Francis Journal Suggester[7]등 해외의 여러 출판사에서 내용 기반의 저널 추천 서비스를 제공하고 있다. 국내에서는 DBpia의 투고저널 추천 기능[8]이 2023년 상반기에 신규 출시되었다. 이처럼 내용 기반 저널 추천은 온라인 서비스로 구현되어 많은 사람들에게 도움을 주고 있는 만큼 사용자 수요 및 유용성 측면에서 어느 정도 입증되었다고 할 수 있다.

내용 기반 추천에 부정적이었던 연구자들은 텍스트를 이용한 저널추천 방식의 정확도에 대해 의문을 제기하였는데, 이는 LLM(Large Language Model)이 적극적으로 사용되기 이전에는 문서의 문맥을 파악하여 좌표 공간에 사영시키는 작업이 매우 어려웠기 때문이다. 이제는 우수한 성능을 보이는 다양한 기초 모델(foundation models)이 배포되어 이를 사용했을 때 성능향상을 기대할 수 있고 적용도 쉬워져 내용 기반 저널추천 방식이 가질 수 있는 장점이 이전보다 확대되었다. 이러한 점을 고려하여 본 연구에서는 SBERT를 이용한 내용 기반 저널추천 시스템을 제안한다.

2.4 혼합된 추천방식

위에 제시된 방법들은 각각의 장단점을 가지고 있다. 협업 필터링의 경우에는 연구자의 다양한 행동(출판, 열람 등) 이력과 다른 연구자들의 이력을 결합하여 추천하기 때문에 이력이 매우 적거나 없는 연구자들의 경우에는 적용하기 어려우며, 새로운 주제로 연구하고자 하는 기존 연구자들에게도 적용이 곤란하다는 문제가 있다. 즉, 연구자의 콜드 스타트 문제[9](Cold-start problem)를 극복하기 어렵다. 네트워크 기반 추천 역시 공동출판 빈도나 인용 이력을 기반으로 추천을

4) https://journalfinder.elsevier.com
5) https://journalsuggester.springer.com
6) https://journalfinder.wiley.com/
7) https://authorservices.taylorandfrancis.com/publishing-your-research/choosing-a-journal/journal-suggester/
8) https://www.dbpia.co.kr/journal/journal-recommend-search
9) 초기 사용자 혹은 아이템의 이력 부재로 인해 추천 시스템의 성능이 저하되는 문제 (Bobadilla et al., 2012).

진행하므로 연구자의 콜드 스타트 문제에 직면하며, 문서의 내용을 반영할 수 없다는 한계점이 있다. 내용 기반 추천의 경우 문서의 내용을 기반으로 추천하기 때문에 내용상 적합한 저널을 추천할 수 있다는 장점이 있다. 그러나, 신규 저널의 경우 출판된 문서의 수가 충분하지 않아 저널의 콜드 스타트 문제가 걸림돌이 될 수 있다. 이처럼 각각의 방식이 가지는 단점들을 극복하고자 여러 방식을 혼합하는 시도도 있다. Pradhan & Pal (2020)은 내용 기반 추천과 네트워크 기반 추천을 혼합하여 네트워크 기반 추천에서 겪을 수 있는 연구자의 콜드 스타트 문제를 해결하면서 내용 기반 추천에서는 고려하기 어려운 논문 간 관계를 이용하여 추천의 질을 높였다. 이외에도 여러 방식의 장점을 취하기 위해 혼합된 방식의 저널추천 방법이 제안되었다(Liu et al., 2022; Pradhan et al., 2020). 혼합된 방식의 저널추천은 콜드 스타트 문제를 완화하고 추천의 질을 향상시키는 데에 도움이 될 수 있지만, 이 역시 완벽한 해결책은 아니다. 왜냐하면, 부족한 데이터로 인해 발생하는 콜드 스타트 문제를 다른 종류의 이력 데이터를 이용하여 완화하므로 이종의 가용 데이터를 보유하고 있지 않다면 적용할 수 없기 때문이다. 이처럼 모든 추천방식에는 장단점이 있으므로 최종 사용자의 요구사항, 가용 데이터 상황, 보유 인프라 등을 종합적으로 고려하여 추천방식을 결정해야 한다.

3. 데이터

본 연구에서는 KISTI에서 제공하는 논문 전문 텍스트 데이터셋(한국과학기술정보원, 2021)을 이용하였다. 이 데이터셋은 KISTI에서 운영하는 국내 학술정보를 위한 개방형 플랫폼 KoreaScience[10]로부터 수집되었다. 이 데이터셋에는 1985년에서 2020년까지 953개의 학술지(학술대회 자료집 포함)에 게재된 481,578개 문서의 제목, 초록, 키워드, 본문 등이 JSON 형태로 수록되어 있다. 데이터셋에서 아래와 같은 조건에 해당되는 건은 모두 제외하였다.

- 문서의 제목, 초록, 키워드 중 하나라도 존재하지 않음
- 학술대회 자료집
- 저널 당 출판논문이 10개 미만

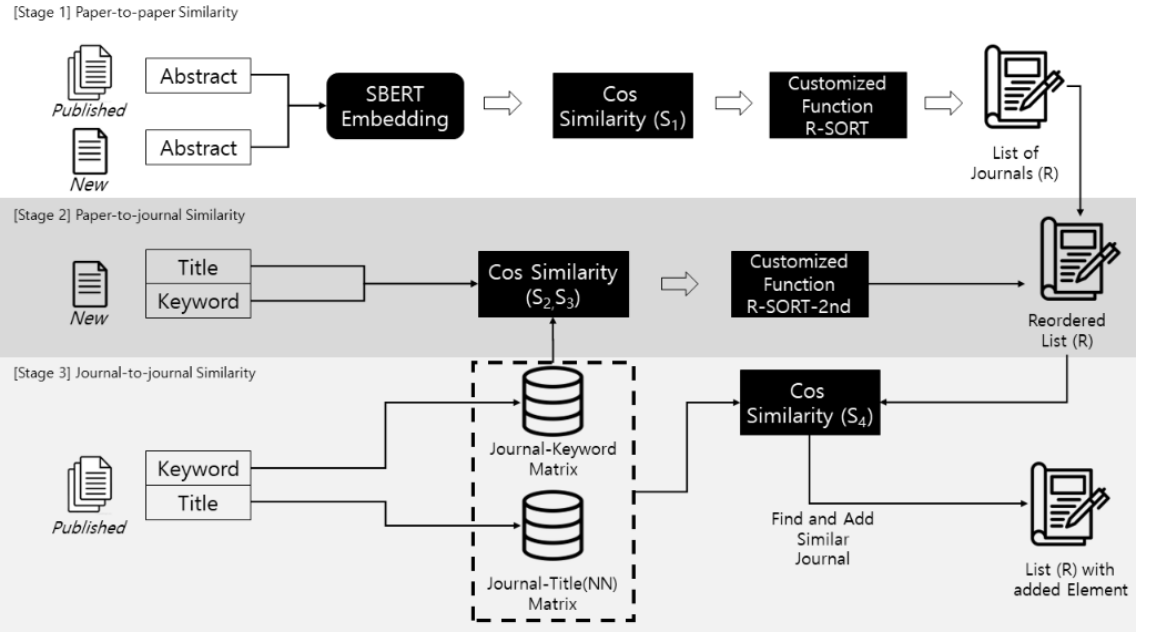
조건을 만족하는 총 268개 저널에서 출판된 103,067개의 논문을 대상으로 연구를 진행하였다.

4. 연구방법

4.1 SBERT(Sentence-BERT)

신진 연구자 및 기존 연구자들의 신규 연구주제에 적합한 저널을 찾을 수 있도록 문서의 내용을 반영할 수 있으면서, 저널의 범위에 큰 영향을 받지 않고 일관된 성능을 낼 수 있는 콘텐츠 기반 추천방식을 제안한다. 본문의 핵심적인 내용을 요약해서 수록하는 초록을 임베딩하여 투고하고자 하는 논문 초록과의 내용 유사도를 계산하는 것이 본 연구에서 제안하는 프레임워크의 핵심이다. 양질의 임베딩을 얻기 위해 KR-SBERT-v40K-KlueNLI-augSTS 모델(박수지, 2021)을 사용하였다. 이 모델은 KR-BERT-v40K(Lee et al., 2020)모델을 기반으로 KLUE-NLI(Park et al., 2021)와 증

10) https://koreascience.kr/



〈Figure 1〉 Suggested Framework

강된 KorSTS(Ham et al., 2020) 데이터셋을 이용하여 미세조정된 Sentence-BERT 모델이다.

Sentence-BERT(SBERT)는 BERT 모델을 샴 네트워크(siamese network)와 트리플렛 네트워크(triplet network)를 이용하여 파인튜닝하는 과정을 거친 것이다. 이때, 각 문장 쌍을 비교하여 유사도를 측정하는 방법을 이용하여 각 문장을 벡터 형태로 변환한다. 이 덕분에 SBERT는 문서의 의미를 보존하면서 벡터 공간에서의 거리를 유사도로 해석할 수 있도록 한다. 따라서, SBERT를 이용하여 문서를 벡터로 변환하면 유사성계산에 적합한 형태로 변환되기 때문에(Reimers & Gurevych, 2019) 기존의 BERT보다 이러한 종류의 태스크에 적합하다고 할 수 있다. 또한, 계산 효율성 측면에서도 SBERT는 기존의 BERT보다 효과적이어서(Lombardo et al., 2022) 긴 문장을 임베딩하기에 알맞다. 특히 최근에는 SBERT 임베딩을 이용한 영문 저널 추천 시스템이 Word2vec, Glove, FastText등 다른 임베딩 방식보다 우수하다는 연구결과도 제시되었다(Gündoğan et al., 2023). 이러한 점들을 고려하였을 때 초록 임베딩에 SBERT를 사용하는 것은 계산효율성 및 성능 측면에서 바람직하다고 할 수 있다.

4.2 코사인 유사도(Cosine Similarity)

코사인 유사도는 간단하고 효과적이며 가장 많이 사용되는 유사도 지표 중 하나이다(Xia et al., 2015). 이는 두 벡터 사잇각의 코사인값으로, 벡터 공간에서 벡터의 스칼라 곱(dot product)을 두 벡터 크기의 곱으로 나눈 값이다(김선경 등, 2020). 두 벡터가 정확히 같은 방향을 향할 때 두 벡터의 사잇각은 0도이며 코사인 유사도는 1이

된다. 반대로 두 벡터가 완전히 반대 방향을 향하면 코사인 유사도가 -1이 되고, 두 벡터가 수직일 때는 0이 된다. 벡터의 길이는 코사인 유사도에 영향을 주지 않으며, 오직 벡터의 방향만 유사도에 영향을 미친다.

4.3 제안된 프레임워크

제안된 프레임워크를 3단계로 나누어 요약하면 <Figure 1>과 같다. 먼저 SBERT를 이용하여 문서 대 문서 유사도를 이용하여 추천할 저널의 목록을 생성한다, 2단계에서는 키워드나 제목이 입력된 경우 문서 대 저널 유사도를 통해 1단계에서 생성된 저널 목록의 순서를 재정렬한다. 제목이나 키워드가 입력되지 않았으면 2단계는 생략한다. 마지막으로 3단계에서는 추천할 저널 목록에서 가장 순위가 높은 저널과 비슷한 유사도를 가진 저널 하나를 저널 대 저널 유사도를 통해 발굴한다.

사용자의 입력에 따라 k개의 저널을 추천할 경우 상세 과정은 다음과 같다. 먼저, 사용자가 입력한 문서와 출판된 모든 문서의 SBERT 임베딩에 대해 코사인 유사도를 계산하고, 이후 해당 문서들이 출판된 저널을 확인한다. 여기에서 특정 기준에 따라 k-1개의 저널을 순차적으로 추천할 리스트에 추가한다. 여기서 문서 간 코사인 유사도(S_1)는 다음과 같은 방식으로 계산된다.

$$S_1 = \frac{SBERT(DA_m) \cdot SBERT(DA_n)}{\| SBERT(DA_m) \| \cdot \| SBERT(DA_n) \|}$$

$SBERT$ = $SBERT$ 모델을 통한 임베딩
DA = 문서 D의 초록
m = 기존 보유 문서
n = 입력받은 문서

유사도 S_1 계산이 끝나면 후보 저널 리스트(R)의 생성을 위해 저널별 점수($R\text{-}SORT_j$)를 구한다. $R\text{-}SORT_j$는 문서 간 유사도 S_1에 지수함수를 적용한 부분과, 일정 S_1 이상의 문서를 다수 보유한 저널에 가중치를 두는 부분으로 나누어진다.

$$f(S_1) = 0.5 + 0.135e^{0.0891 \times S_1}$$

$$R-SORT_j = \frac{\sum_{i=1}^{l} f(S_{1_i}) + \sum_{i=1}^{l} c}{l}$$

j = 기존 보유 저널
l = 저널에 출판된 문서 수
c = 저널 j에서 S_1이 70% 이상인 문서 수

먼저 문서 간 유사도에 지수함수를 적용한 $f(S_1)$의 값을 계산한다. $f(S_1)$은 <Figure 5>와 같이 유사도가 70%일 때 70, 유사도가 100%일 때 1000의 값을 가지는 지수함수이다. 도메인, 임베딩 벡터, 작업의 종류에 따라서 유사성을 판단하는 코사인 유사도의 기준값은 달라진다(Orkphol & Yang, 2019). 이에, 연구자들은 여러 문서들의 내용과 코사인 유사도를 살펴본 후 70% 이상의 코사인 유사도를 가지는 문서를 관련 문서로 정의하고, 유사도 70% 이상의 문서에 더욱 비중을 두도록 가중치를 적용하였다. 즉, 유사도 S_1에 지수함수를 적용한 것은 유사도가 70%보다 큰 문서에는 더 큰 가중치를 주고 유사도가 70%보다 작은 문서에는 더욱 작은 가중치를 주는 방식으로 추천 저널을 결정하기 위함이다. 이를 통해 유사도가 매우 높은 문서가 포함된 저널은 저널에 출판된 문서의 수가 다소 적더라도 추천 리스트에 포함되게 된다. 이러한 비례적 변화 관계를 모델링하기 위해 지수함수는 기초 및 응용과학 연구에서 널리 사용된다(Goldstein et al., 2006).

이후 추천 저널을 결정하기 위해 <Figure 2>와 같이 $f(S_1)$을 저널별로 합한 값과, S_1이 일정 유사도

(70%) 이상을 가지는 문서 수(c)를 해당 저널의 문서 개수(l)만큼 저널별로 더하여 평균한다. 저널별로 R-SORT$_j$가 계산되면 이 값의 순서대로 저널을 정렬하고 여기에서 k-1개의 저널을 후보 저널 리스트(R)에 추가하여 1단계를 마무리한다.

	Document	S_1	$f(S_1)$	...	c	R-SORT$_j$
J_1	D_1	0.90	410.6	...	1	(410.6+1.4+2)/2 =414/2 =207
	D_2	0.21	1.4	...	1	
...	...	...	...	...	...	...
J_j	D_3	0.70	70	...	3	(70+70+70+9)/3 =219/3 =73
	D_4	0.70	70	...	3	
	D_5	0.70	70	...	3	

〈Figure 2〉 Calculation of R-SORTj

R-SORT$_j$에 저널 가중치를 적용하지 않고 $f(S_1)$만으로 추천 리스트를 결정할 경우 지수함수의 특성 때문에 적정한 유사도를 보이는 문서가 다수 포함된 저널이 저평가된다. 따라서, 이러한 문제를 방지하고자 일정 유사도 이상의 문서를 많이 포함한 저널 역시 추천 리스트에 포함되도록 유도하여 추천의 안정성을 확보하고자 하였다. 이와 같은 방식으로 유사도가 매우 높은 문서가 포함된 저널은 발행문서가 소수이더라도 추천 리스트에 포함되게 함과 동시에 전반적으로 유사하다고 판단되는 저널도 리스트에 포함하도록 하여 추천 결과의 안정성을 확보하였다. 두 가지 측면을 고려하여 리스트를 생성하도록 제안하는 배경은 5.2절에서 다시 설명할 것이다.

초록 외 키워드나 제목이 입력된 경우에는 2단계에서 다음과 같이 문서 대 저널 유사도 비교 과정을 진행한다. 키워드가 입력된 경우 저널에 출판된 문서들의 키워드와 일치도를 이용해 리스트의 정렬상태를 변경하여 추천 결과를 개선한다. 이를 위해 저널별로 출판된 논문들의 키워드를 모두 모아서 Journal-Keyword matrix(JKM)를 생성한다. 여기서 JKM은 <Figure 3>와 같이 기존에 출판된 문서의 term frequency를 저널별로 더하여 생성한다.

Document-Keyword Matrix

	Keyword / Document	K_1	K_2	...	K_k
J_1	D_1	1	3	...	1
	D_2	3	0	...	1
...	...	...	...	...	...
J_j	D_m	4	2	...	0

⇨

Journal-Keyword Matrix

Keyword / Journal	K_1	K_2	...	K_k
J_1	4	3	...	2
J_2	0	1	...	0
...	...	...	...	...
J_j	5	3	...	1

〈Figure 3〉 Journal-Keyword Matrix

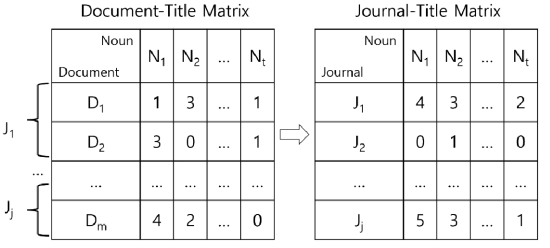
이 행렬과 신규 논문의 키워드 간 유사도를 구하기 위해 마찬가지로 코사인 유사도를 이용한다. 생성된 JKM을 이용한 저널 대 문서의 키워드 유사도(S$_2$)는 아래와 같이 산출한다.

$$S_2 = \frac{JKM_j \cdot DKM_n}{\| JKM_j \| \cdot \| DKM_n \|}$$

$JKM = Journal - Keyword\ Matrix$
$DKM = Document - Keyword\ Matrix$
j = 기존 보유 저널
n = 입력받은 문서

마찬가지로 제목이 입력된 경우 신규 논문과 각 저널에 출판된 논문 제목과의 일치도를 계산하여 추천 결과를 개선한다. 구체적으로는 각 저널에 출판된 논문들의 제목에서 명사(일반명사 및 고유명사)를 추출하여 Journal-Title matrix(JTM)를 생성하고, 이 행렬과 신규 문서의 제목에서 추출한 명사들과의 유사도를 구한다. 이는 이전

연구에서 유사도 계산을 통한 저널추천 및 성능 개선을 위해 명사를 이용하였다는 점을 고려한 것이다 (Kang et al., 2015; 이호엽 등, 2015; 유영선, 2015).



〈Figure 4〉 Journal-Title Matrix

생성된 JTM을 이용한 저널 대 문서의 제목 유사도(S_3)는 다음과 같은 방식으로 계산한다.

$$S_3 = \frac{JTM_j \cdot DTM_n}{\| JTM_j \| \cdot \| DTM_n \|}$$

$JTM = Journal-Title\ Matrix$
$DTM = Document-Title\ Matrix$
j = 기존 보유 저널
n = 입력받은 문서

이처럼 제목 혹은 키워드가 입력된 경우 문서 대 저널 비교를 통해 계산된 유사도 S_2 및 S_3을 이용하여 추천할 저널 리스트(R)의 정렬상태를 개선한다.

$$R-SORT-2nd_j = S_2 + S_3$$

이때, 키워드만 입력되었다면 S_3=0이며, 제목만 입력되었다면 S_2=0이다. 최종적으로 R-SORT-2nd$_j$가 높은 저널 순으로 추천 저널 리스트(R)에 속한 저널을 재정렬한다. 키워드와 제목이 입력되지 않으면 S_2와 S_3 계산이 불가능하므로 2단계를 진행할 수 없으므로 생략하고 다음 3단계를 진행한다.

3단계에서는 추천 저널 리스트에서 상위 1개의 저널(R_1)을 선택한 후 저널 대 저널 비교를 통해 이 저널과 가장 유사한 제목과 키워드 분포를 보이는 저널을 추가로 추천할 리스트에 포함시킨다. 여기서 저널 간 유사도(S_4)는 JKM_j과 R_1과의 유사도, 그리고 JTM_j과 R_1과의 유사도의 합이다. 이때 추가되는 저널이 기존 추천할 리스트에 있는 저널과 동일한 저널일 경우에는 새로운 저널이 추가될 때까지 차순위의 유사도를 보이는 저널 1개를 추가하도록 시도한다. 물론 필요에 따라 이 방법을 이용하여 몇 개의 저널을 더 추가할 수도 있으며, 일정 유사도 이상의 저널들만 추가하는 방식도 가능하다. S_4 계산 시 R_1만 사용하지 않고 R_2나 R_3등 차순위 저널과의 유사도를 동시에 고려하는 방식도 생각할 수 있다.

$$S_4 = \frac{JKM_j \cdot JKM_{R_1}}{\| JKM_j \| \cdot \| JKM_{R_1} \|} + \frac{JTM_j \cdot JTM_{R_1}}{\| JTM_j \| \cdot \| JTM_{R_1} \|}$$

$JKM = Journal-Keyword\ Matrix$
$JTM = Journal-Title\ Matrix$
j = 기존 보유 저널
R_1 = 리스트 R의 1번째 원소

제안된 프레임워크의 장점은 초록만 입력해도 유사한 논문 검색과 후보 저널 추천이 가능하다는 점이다. 또한, 선택적으로 제목이나 키워드를 입력할 경우 추천의 질을 개선할 수 있다. 마지막 단계에서 저널 대 저널 유사도를 비교하여 문서 대 문서 유사도에서 발견되지 않았던 저널을 추천 리스트에 포함하여 추천의 다양성도 높일 수 있다.

5. 실험

5.1 추천 정확도 평가지표

평가지표로는Top-k 정확도와MRR(Mean Reciprocal Rank)을 사용한다. Top-k 정확도는 분류 모델이 생성한 가장 높은 k개의 예측 중에서 정답이 존재하는 경우 올바른 예측으로 간주하는 방식이다. 다시 말해, Top-1 정확도의 경우 우리가 흔히 사용하는 정확도 개념과 동일하고, Top-3 정확도는 모델의 상위 3개 예측 중 하나가 정답일 때 정답으로 간주하는 것이다. 일반적으로 k가 커질수록 모델이 출력하는 예측의 개수가 많아지므로 Top-k 정확도는 높아진다.

Top-k는 추천 리스트에 등장한 순서와 관계없이 평가하나, 추천된 리스트의 순서 역시 추천의 질에 영향을 미친다. MRR은 정답 아이템이 리스트 중 몇 번째에 위치하는지를 고려하여 평가할 수 있는 지표이다. 추천된 리스트 R에서 정답이 K_u면 RR(Reciprocal Rank)은 $1/k_u$가 된다. 예를 들어, 추천 리스트에서 3번째에 정답이 있다면 RR은 1/3이 되고 4번째에 정답이 있다면 1/4이 된다. 각각의 추천 리스트별로 계산된 RR에 대해서 평균을 내면 MRR이 된다. 이렇게 계산된 MRR은 0에서 1까지의 값을 가지며 1에 가까울수록 추천의 질이 높다고 할 수 있다.

$$MRR = \frac{1}{|R|}\sum_{u \in R}\frac{1}{k_u}$$

5.2 추천 정확도 평가결과

주어진 데이터셋을 평가하기 위해 저널별 9:1 비율로 계층적 분할하였다. 앞서 언급된 것처럼 Top-k 정확도 및 MRR을 통해 평가를 진행하였다. 추천 리스트에 포함된 추천 저널의 수를 3개, 5개, 10개, 15개, 20개로 변화시켜가면서 Top-k 정확도와 MRR을 산출하였다. 여기서 Top-k 정확도는 micro 정확도와 macro 정확도 두 가지로 측정되었다. 개별 케이스(논문)에 동일한 가중치를 적용한 것이 micro 정확도이며, 각각의 클래스(저널)별로 micro 정확도를 먼저 계산한 후 클래스별 정확도의 평균을 계산하는 것이 macro 정확도이다. 즉 macro 정확도는 각 저널에 동등하게 중요도를 부여하므로 소수 클래스(출판된 논문의 개수가 적은 저널)에 대한 예측 결과값이 좋지 못한 경우 micro 정확도보다 낮은 값을 기록하게 된다.

먼저 키워드와 제목 입력 없이 초록만으로 1단계만 거쳐 k개의 저널을 추천할 경우의 평가결과는 <Table 1>과 같다. 초록만 입력받아 추천하였음에도 Top-10 micro 정확도 75.4%를 기록하였다.

〈Table 1〉 Evaluation Metrics
Input: Abstract (Stage 1)

	Top3	Top5	Top10	Top15	Top20
Micro Acc.	0.520	0.636	0.754	0.817	0.855
Macro Acc.	0.433	0.554	0.680	0.753	0.807
MRR	0.388	0.414	0.431	0.436	0.438

<Table 2>~<Table 4>에는 키워드나 제목이 입력되고 제안된 프레임워크 3단계를 모두 이행한 경우의 평가지표가 나타나 있다. 키워드와 제목이 모두 입력된 <Table 4>의 Top-10 micro 정확도는 76.6%를 기록하였다. 이는 추천 시스템으로부터 제시된 10개의 추천 저널 내에 특정 논문이 실제로 게재된 저널이 포함될 확률이 76.6%를 의미하며, 수용 가능한 수준의 결과라고 볼 수 있다.

〈Table 2〉 Evaluation Metrics
Input: Abstract, Keyword (Stage 1~3)

	Top3	Top5	Top10	Top15	Top20
Micro Acc.	0.495	0.632	0.765	0.828	0.865
Macro Acc.	0.409	0.541	0.688	0.760	0.814
MRR	0.386	0.431	0.442	0.436	0.429

〈Table 3〉 Evaluation Metrics
Input: Abstract, Title (Stage 1~3)

	Top3	Top5	Top10	Top15	Top20
Micro Acc.	0.494	0.631	0.764	0.826	0.862
Macro Acc.	0.409	0.539	0.685	0.758	0.810
MRR	0.386	0.425	0.429	0.416	0.406

〈Table 4〉 Evaluation Metrics
Input: Abstract, Title, Keyword (Stage 1~3)

	Top3	Top5	Top10	Top15	Top20
Micro Acc.	0.495	0.632	0.766	0.827	0.864
Macro Acc.	0.409	0.541	0.689	0.760	0.814
MRR	0.391	0.441	0.463	0.459	0.456

저널별 정확도의 평균인 macro 정확도의 값이 전체 정확도 평균인 micro 정확도 값보다 공통적으로 소폭 낮게 나타나는데, 게재된 논문이 많은 저널의 정확도는 높게 나오고 게재된 논문이 적은 저널들의 정확도는 낮게 나오고 있기 때문이다. 이러한 현상은 앞서 언급되었던 저널의 콜드스타트 문제로 인해 유발되는데, 특히 저널의 범위가 넓은 초기저널의 경우 낮은 micro 정확도를 기록하기 쉽다. 다른 주목할만한 점은 3단계인 저널 대 저널 유사도로 유사저널을 추가하는 단계가 추천 정확도를 저해하기도 하고 증대시키기도 한다는 점이다. 아래 <Table 5>~<Table 7>은 저널 대 저널 유사도를 통해 마지막 요소를 추가하는 3단계 없이 1단계에서 k개의 저널을 선택한 후 2단계에서 리스트의 재정렬을 진행한 경우의 정확도이다. <Table 2>~<Table 4>의 Top-3 및 Top-5 정확도를 <Table 5>~<Table 7>의 Top-3 및 Top-5 정확도와 비교하였을 때 3단계를 진행하지 않은 경우(후자)의 정확도가 더 높은 것을 확인할 수 있다. 반대로 k가 높아지면 3단계를 진행한 경우의 정확도가 높아지는 것을 확인할 수 있다. 이는 충분히 많은 수의 저널을 추천할 경우 저널 대 저널 유사도를 활용하는 것이 추천의 질을 저하시키지 않으면서 추천방식의 다양성을 증대시키는 방법으로서 유효하다는 것을 보여준다.

〈Table 5〉 Evaluation Metrics
Input: Abstract, Keyword (Stage 1~2)

	Top3	Top5	Top10	Top15	Top20
Micro Acc.	0.520	0.636	0.754	0.817	0.855
Macro Acc.	0.433	0.554	0.680	0.753	0.807
MRR	0.402	0.433	0.438	0.433	0.426

〈Table 6〉 Evaluation Metrics
Input: Abstract, Title (Stage 1~2)

	Top3	Top5	Top10	Top15	Top20
Micro Acc.	0.520	0.636	0.754	0.817	0.855
Macro Acc.	0.433	0.554	0.680	0.753	0.807
MRR	0.398	0.427	0.424	0.412	0.403

〈Table 7〉 Evaluation Metrics
Input: Abstract, Title, Keyword (Stage 1~2)

	Top3	Top5	Top10	Top15	Top20
Micro Acc.	0.520	0.636	0.754	0.817	0.855
Macro Acc.	0.433	0.554	0.680	0.753	0.807
MRR	0.410	0.447	0.460	0.457	0.454

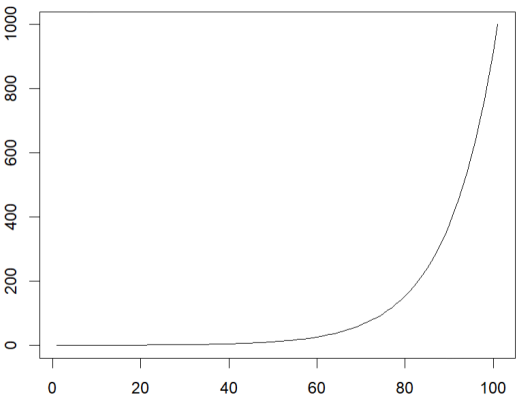
<Table 1>과 <Table 5>~<Table 7>를 비교해 보면 키워드와 제목이 입력되어 리스트가 재정렬되는 경우 추천의 질이 높아지는 것을 알 수 있다. 키워드와 제목을 이용하여 추천 저널 리스트를 재정렬하는 2단계가 유의미한 작업이라는 점을 확인할 수 있는 결과이다. 키워드와 제목이 모두 입력된 경우 평가지표가 가장 크게 개선되었으며, 제목 내의 명사와 키워드 중에서는 키워드의 중요성이 더 큰 것으로 나타났다. 저널추천 서비스를 제공하는 사이트에서 대부분 키워드나 제목을 필수로 기재하도록 설정해놓는 경향이 있는데 이러한 이유 때문으로 추측할 수 있다. 한편, <Table 2>~<Table 7> 모두 k가 증가할 때 MRR 값이 함께 증가하다가 k가 15 이상으로 커지면 오히려 줄어드는 것을 볼 수 있는데, 이는 1단계에서 추천에 적절한 저널이 충분히 포함된 상태로 리스트가 생성되지 않으면, 키워드와 제목만으로 저널을 재정렬하는 것이 추천 결과를 악화시킬 수 있다는 것이다. 즉, 보유 저널과 문서 상태에 따라 k를 적절한 수로 조절해야 추천의 질이 저해되지 않은 상태에서 올바른 저널을 추천할 수 있다.

마지막으로 제안된 프레임워크 1단계에서 사용된 $f(S_1)$ 및 R-SORT$_j$ 형태에 따른 정확도 평가 결과를 추가로 비교하여 4.3절에서 제안된 R-SORT$_j$의 형태의 장점을 설명한다. <Table 8>은 $f(S_1)$과 R-SORT$_j$를 아래와 같이 구성하였을 때 결과이다.

$$f(S_1) = 0.5 + 0.135e^{0.0891 \times S_1}$$

$$R-SORT_j = \frac{\sum_{i=1}^{l} f(S_{1_i})}{l}$$

j = 기존 보유 저널
l = 저널에 출판된 문서 수



〈Figure 5〉 $f(S_1)$: Exponential function

〈Table 8〉 Evaluation Metrics*
Input: Abstract, Title, Keyword (Stage 1~3)

	Top3	Top5	Top10	Top15	Top20
Micro Acc.	0.380	0.477	0.593	0.662	0.715
Macro Acc.	0.481	0.594	0.708	0.776	0.810
MRR	0.292	0.335	0.370	0.387	0.399

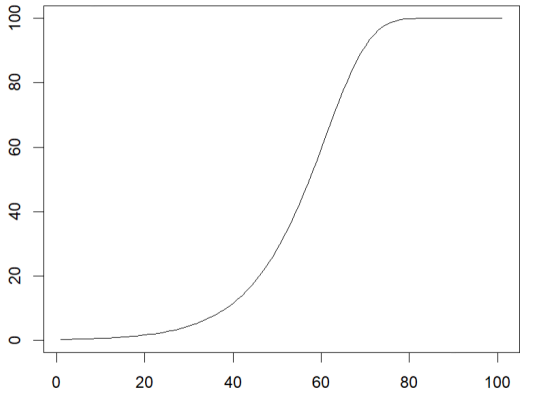
* Experiment with modified $f(S_1)$ in <Figure 5>

<Table 9>는 R-SORT$_j$를 일정 유사도 이상의 문서를 많이 소유한 저널에 유리하도록 $f(S_1)$을 다음과 같이 곰페르츠(Gompertz) 함수 형태로 변경했을 때의 결과이다. 여기에서 $f(S_1)$은 <Figure 6>과 같이 유사도 70% 근처에서 출력 100에 수렴하는 S자 형태의 함수이다.

$$f(S_1) = 100 \times 1 - e^{-e^{0.1 \times (S_1 - 60)}}$$

$$R-SORT_j = \frac{\sum_{i=1}^{l} f(S_{1_i})}{l}$$

j = 기존 보유 저널
l = 저널에 출판된 문서 수



〈Figure 6〉 $f(S_1)$: Gompertz function

〈Table 9〉 Evaluation Metrics*
Input: Abstract, Title, Keyword (Stage 1~3)

	Top3	Top5	Top10	Top15	Top20
Micro Acc.	0.330	0.427	0.554	0.619	0.672
Macro Acc.	0.420	0.535	0.669	0.740	0.781
MRR	0.249	0.297	0.345	0.364	0.377

* Experiment with modified $f(S_1)$ in <Figure 6>

<Table 8> 및 <Table 9>에 적용된 $f(S_1)$이 전혀 다른 형태임에도 불구하고 우수한 정확도를 기록하는 것을 확인할 수 있다. 이러한 결과를 토대로 두 가지 방식을 결합하기 위해 더 높은 정확도를 보이는 지수함수 형태의 $f(S_1)$을 적용하면서, 70% 이상의 유사도를 보이는 문서들이 많은 저널에 가중치를 두기 위해 R-SORT$_j$에 $\sum c$ 부분을 추가하여 4.3절의 형태로 제안한 것이다. 기계학습에서 앙상블(ensemble) 방법 적용 시 추론방식이 다른 우수한 모델을 둘 이상 결합할 경우 성능향상을 가져온다(Sagi & Rokach, 2018)는 점에 착안하여, 서로 다른 방식의 추론을 결합할 수 있도록 R-SORT$_j$를 구성하였다. <Table 8>~<Table 9>에 기록된 정확도보다 두 가지 방식을 결합하여 제안된 프레임워크를 이용한 결과인 <Table 4>에 나타난 정확도가 전반적으로 우수한 것을 확인할 수 있다.

5.3 사용자 평가결과

5.2절에서 제시된 Top-k 정확도는 충분히 수용 가능한 수준이지만, 사용자 관점에서의 추천 적합도를 평가하기 위해서는 사용자의 피드백도 필요하다. 왜냐하면, 실제 논문이 게재된 저널이 추천목록에 포함되지 않았지만 다른 적합한 저널들이 추천된 경우 성능이 과소평가 될 수 있기 때문이다. 특히 주제가 여러 저널의 범위에 부합하는 주제이거나, 신생 저널에 게재된 논문을 테스트하는 경우 지표와 실제 사용자 체감 성능의 괴리가 있을 수 있다. 이 때문에 추천 결과를 실제 사용자로부터 평가받는 작업도 필요하다. 평가에 사용했던 테스트셋을 분야별로 나누고 최근 1년 내 연구 활동 및 저널 투고 경험이 있는 7명에게 추천 결과의 평가를 요청하였다. 평가에 참여한 참여자들의 프로파일은 <Table 10>과 같다.

〈Table 10〉 Assessment participants

No.	Profile	Research Field
1	연세대학교 경영학 박사졸업	경영정보시스템
2	서울연구원 연구원	교통/ 도시행정
3	한양대학교 기술경영학 박사수료	기술경영/ 기술혁신
4	한양대학교 기술경영학 박사수료	인공지능/ 지능형시스템
5	LG CNS 선임연구원	인공지능/ 지능형시스템
6	삼성디스플레이 책임연구원	기계공학
7	LG AI Research 선임연구원	인공지능/ 지능형시스템

평가 참여자들에게 관련 연구 분야에 해당하는 저널에 게재된 논문의 초록, 제목, 5개의 추천 결과를 제공하였다. 관련 연구 분야에 부합하는 저널에 실제로 게재된 논문에 대해서만 평가하도록 하였으나, 판단이 어려운 논문의 추천 결과에 대해서는 응답하지 않을 수 있도록 하였다. 평가 참여자들은 논문의 제목과 초록을 읽고 제시된 5개의 추천 저널이 투고에 적합한지 아닌지를 판단한다. 즉, 평가자들은 1개의 논문 당 5개의 저널추천 결과에 대해 저널의 목적과 범위를 고려하여 해당 논문이 투고에 적합한지 혹은 투고에 부적합한지 구분하도록 요청된다. 정확도 평가 시 사용된 지표는 5.2절과 동일하다. 다만, 5.2절에서는 실제 게재된 저널이 추천 리스트에 올바르게 등장해야 정답으로 처리하였으나, 여기서는 사용자의 판단으로 적합하다고 인정된 저널이 리스트에 존재하면 정답처리를 하게 된다는 점이 차이점이다.

〈Table 11〉 User Evaluation

No.	Number of Papers Evaluated	Top5 Acc.	MRR
1	65	0.938	0.769
2	58	1.000	0.991
3	66	1.000	0.985
4	93	0.978	0.888
5	99	0.788	0.623
6	146	0.979	0.889
7	99	0.960	0.856

7명의 평가자는 총 626개의 문서를 평가하였으며, 평가 참여자들이 응답한 결과에 따라 산출한 Top-5 micro 정확도는 0.946, Top-5 macro 정확도는 0.949, MRR은 0.850으로 나타났다. 실제 사용자들의 평가결과는 단순히 실제 게재된 저널을 올바르게 추천 리스트에 포함하였는지를 나타낸 5.2절의 <Table 4>보다 현저히 높은 것을 확인할 수 있다. 이는 추천된 저널 리스트에 해당 논문이 실제 게재된 저널이 포함되어 있지 않더라도, 투고가 가능한 다른 저널을 추천하고 있음을 보여준다. 결과를 평가자별로 세분화하여 기록한 내용이 <Table 11>에 제시되어 있다.

평가자들은 적게는 인당 58개에서 많게는 146개 논문과 추천 리스트를 비교하여 리스트에 제시된 저널이 투고에 적합한지 평가하였다. 평가 과정 중 더 엄격하게 적합 여부를 판단하는 평가자도 있었고, 보다 허용적으로 판단한 평가자도 있었다. 평가자들의 Top-5 정확도는 0.788~1.000, MRR은 0.623~0.991을 기록하였다. 평가결과로 미루어볼 때 사용자 관점에서의 추천 결과 유효성은 충분하다고 볼 수 있다. 하지만 위의 결과로는 연구 분야별 추천의 품질 차이, 평가의 개인차, 개선 방향 등을 정확히 알기 어렵다. 고도화를 위해서는 실제로 웹에서 서비스를 제공하면서 많은 사용자로부터 대규모 피드백을 받아 추천 시스템을 개선하는 것이 이상적이다.

5.4 본고의 추천 결과

본 논문의 초록, 제목, 키워드를 입력하고 제안된 프레임워크를 통해 3개의 후보 저널을 추천받은 결과 <Table 12>와 같은 결과가 도출되었다.

〈Table 12〉 Result of recommendation for this article

순위	학술지 이름	ISSN	연구분야
1	지능정보연구	2288-4866	공학> 산업공학
2	한국전자거래학회지	2288-3908	공학> 컴퓨터학
3	인터넷정보학회논문지	1598-0170	공학> 컴퓨터학

3순위로 추천된 인터넷정보학회눈문지(Journal of Internet Computing and Services, JICS)는 인터넷과 정보통신기술을 융합한 통신, 보안, 미디어, 소프트웨어, 지능형 시스템, IT정책 및 서비스 등에 대한 연구 성과를 발표하는 학술지이다. 제출 트랙 목록에는 정보 검색 및 필터링, AI 및 지능형 시스템 등이 포함되어 있다. 저널 범위와 제출 트랙을 고려하였을 때 추천이 적합하다고 볼 수 있다.

2순위로 추천된 한국전자거래학회지(Journal of Society for e-Business Studies, JSEBS)는 e-Business의 모델 및 정책에 대한 연구, e-Business를 통한 혁신에 대한 설계 또는 적용, e-Business 인프라를 위한 컴퓨터 및 정보 기술에 대해 다루고 있다. 저널의 범위를 고려하였을 때 두 번째 추천 결과 역시 적합하다고 볼 수 있다.

1순위로 추천된 지능정보연구(Journal of Intelligence and Information Systems, JIIS)는 정보기술, 시스템 및 그 응용에 대한 새로운 결과 및 독창적인 아이디어를 다루고 있으며, 정보 시스템, 비즈니스 인텔리전스 모델, 시스템 개발 방법, 데이터 분석, 기계학습, 인공지능 등의 주제를 포함하고 있다. 저널의 범위를 고려하였을 때 첫 번째 추천 저널 역시 매우 적합하다고 볼 수 있다. 추천 결과를 바탕으로 본고를 지능정보연구에 투고하였다.

6. 결과 및 논의

본 연구에서는 논문을 투고하고자 하는 연구자가 작성한 초록을 기반으로 투고할 수 있는 저널을 추천하는 시스템의 프레임워크를 제안하였다. 문서 대 문서 유사도, 저널 대 문서 유사도, 저널 대 저널 유사도를 반영하여 추천의 질을 높이도록 고안하였으며, 이 과정에서 SBERT 및 코사인 유사도 개념을 이용하였다. 제안된 프레임워크를 이용하여 평가한 결과 Top-10 정확도 76.6%의 정확도를 기록하였다. 다양한 시나리오를 비교한 결과 키워드나 제목을 추가로 입력받는 경우 문서 대 저널 유사도를 이용하여 추천의 질을 개선할 수 있다는 점을 확인하였으며, 저널 대 저널 유사도를 이용하여 추천의 질을 저하시키지 않으면서 추천방식의 다양성을 확보할 수 있음도 보였다.

추천 시스템의 추천 정확도와 더불어 해당 시스템이 무리 없이 서비스로 구현이 가능할 정도의 자원요구량과 구현 용이성을 갖추었는지 확인하는 것은 중요하다. 나아가, 사용자 관점에서 지속적으로 서비스를 개선하고 사용자 피드백을 반영해 사용자들이 원하는 시스템을 구축하고 사용자에게 유용한 서비스의 형태로 제공하는 것이 실제 서비스 운영에 있어서 중요하다. 제안된 프레임워크는 공급자의 상황 (저널 및 문서 보유 상황, 인프라 활용 수준 등)에 따라서 유연하게 변경 및 적용을 할 수 있어 실제 적용 가능성이 높은 것이 장점이다. 예를 들어 문서 대 문서 유사도(S_1) 계산 후 추천 저널을 결정하기 위한 함수 $f(S_1)$를 변경하거나, 3단계에서 저널 대 저널 유사도를 통해 추천할 저널의 수를 늘릴 수도 있다. 유사도 계산을 위해 사용된 코사인 유사도를 다른 유사도 지표나 거리 기반 지표로 변경하여 사용할 수도 있다. 학술 문서에 특화된 기초모델을 사용하여 임베딩을 변경할 수도 있다. BERT 등의 사전학습 언어모델은 학습에 사용된 말뭉치의 도메인 등 데이터 특성에 영향을 받으므로, 해당 도메인의 말뭉치를 수집 및 활용하여 언어모델을 처음부터 학습시키기도 한다 (김동규 등, 2022). 이처럼 학술문서 말뭉치를 통해 사전학습 모델을 생성하여 임베딩을 생성한다면 추가적인 성능 향상을 기대할 수 있다. 이

외에도 다중 벡터 임베딩과 같은 임베딩 방식을 적용하여 텍스트 내용 기반 추천의 품질을 향상시킬 수도 있을 것이다(박종인 & 김남규, 2019). 이러한 요소변경은 공급자 상황에 맞춘 초기 서비스 구현, 사용자 피드백 반영, 추천의 질 향상에 도움을 줄 수 있다.

제안된 프레임워크가 가지는 약점 중 하나는 주기적으로 JKM 및 JTM을 업데이트해야 한다는 점이다. 그러나 최근 증가된 컴퓨팅 파워와 스토리지를 활용하고, 행렬을 희소행렬 형태로 보관한다면 요구되는 스토리지 자원을 크게 감소시킬 수 있으므로 큰 문제가 되지는 않는다(Goharian et al., 2001). 한편, 내용 기반 저널 추천 시스템이 가지는 구조적인 한계점인 저널의 콜드 스타트 문제를 극복하기 위한 방안이 논의될 필요가 있다. 특히 출판문서의 수가 아직 없거나 아주 적은 신규 저널들은 문서 간 유사도는 물론 문서 대 저널 유사도, 저널 대 저널 유사도에서도 낮은 점수를 기록할 가능성이 높아 최종 추천목록에 등장하기 어려울 수 있다. 이러한 문제를 극복하기 위해서는 추천목록에 등장하기 어려운 신규 저널을 홍보할 수 있는 공간을 추가로 마련하는 등 별도의 조치가 필요하다.

학술지의 성격 변화에 대해서도 생각해 볼 필요가 있다. 학술지 역시 시간이 지나면서 변화를 겪는다. 예를 들어 학술지의 범위가 좁아질 수도 있고 반대로 넓어질 수도 있다. 이러한 변화에 대응하기 위해서 저널별 CFP(Call For Paper)나, 저널의 투고규정 등 논문이 출판되기 전에 학술지의 변화 방향을 미리 알 수 있는 문서들을 활용한다면 더욱 우수한 추천 시스템 구축이 가능할 것으로 사료된다.

11) https://scholar.kyobobook.co.kr

12) https://www.kci.go.kr/

본 연구에서는 국문 저널추천을 위한 프레임워크를 제안하고, 프레임워크의 각 단계가 가져오는 추천 성능의 향상에 대해 실증적으로 제시하였다. 그러나, 프레임워크에 이용된 각 요소들이 추천에 최적이라는 보장은 없으므로 추후 개선 및 고도화가 필요하다고 하겠다. 이 작업은 추후 과제로 남겨둔다.

한국어를 대상으로 하는 저널추천 연구는 지금까지 활발히 이루어지지 않았지만, 이 연구를 시작으로 다양한 연구들이 이루어졌으면 하는 바람이다. 더불어 실무적으로도 쉽게 적용이 가능한 방법이기 때문에 학술저널 열람 서비스를 제공하고 있는 교보문고 스콜라[11], KCI[12] 등의 플랫폼에서 제안된 프레임워크를 응용하여 저널추천 서비스를 제공하는 것을 고려해볼 수 있을 것이다. 이러한 서비스는 이용자 편익 증대에 도움이 됨은 물론이고 사이트 방문을 유도할 수 있는 방안 중 하나로 기능할 수 있을 것이다. 마지막으로 연구를 위한 대규모 학술 데이터 세트가 앞으로도 지속적으로 구축되어 대중에게 공개되기를 기대한다.

참고문헌(References)

김동규, 이동욱, 박장원, 오성우, 권성준, 이인용, & 최동원. (2022). KB-BERT: 금융 특화 한국어 사전학습 언어모델과 그 응용. *지능정보연구*, 28(2), 191-206.

김선경, 박지수, 손진곤. (2020). 유사도 통합에 관한 연구. *한국정보처리학회 학술대회논문집*, 27(2), 53-56.

박수지. (2021). Fusion models for news quality prediction (국내박사학위논문). 서울대학교 대학원, 서울.

박종인, 김남규. (2019). 복합 문서의 의미적 분해를 통한 다중 벡터 문서 임베딩 방법론. *지능정보연구*, 25(3), 19-41.

손연빈, 장태우, 최예림. (2019). 연구자의 논문 게재 이력을 고려한 저널 결정 요인별 중요도 학습 기반의 저널 추천 방법론. *인터넷정보학회논문지*, 20(4), 73-79.

손지은, 김성범, 김현중, 조성준. (2015). 추천 시스템 기법 연구동향 분석. *대한산업공학회지*, 41(2), 185-208.

신은자. (2001). 전자저널의 아카이빙에 관한 연구. *정보관리학회지*, 18(3), 139-158.

유영선 (2015). 딥러닝 알고리즘을 이용한 저널 추천 방법론 (국내석사학위논문). 연세대학교 공학대학원, 서울.

이호엽, 윤휘건, 김창욱. (2015). 딥러닝을 이용한 저널 추천 시스템. *대한산업공학회 추계학술대회 논문집*, 1247-1267.

최인복, 이재동. (2009). 이웃크기를 이용한 사용자기반과 아이템기반 협업여과의 결합예측기법. *정보처리학회논문지: 소프트웨어 및 데이터 공학*, 16(1), 55-62.

최슬비, 곽기영, 안현철. (2016). 사용자 간 신뢰관계 네트워크 분석을 활용한 협업 알고리즘의 예측 정확도 개선. *지능정보연구*, 22(3), 113-127.

한국과학기술정보연구원. (2021). 국내 논문 전문 텍스트 데이터셋 (Version 1.0) [Data set]. 한국과학기술정보연구원. https://doi.org/10.23057/38.

Alhoori, H., & Furuta, R. (2017). Recommendation of scholarly venues based on dynamic user interests. *Journal of Informetrics*, 11(2), 553-563.

Bobadilla, J., Ortega, F., Hernando, A., & Bernal, J. (2012). A collaborative filtering approach to mitigate the new user cold start problem. *Knowledge-based systems*, 26, 225-238.

Chen, Z., Xia, F., Jiang, H., Liu, H., & Zhang, J. (2015). AVER: Random walk based academic venue recommendation. *In Proceedings of the 24th international conference on World Wide Web,* 579-584.

Feng, X., Zhang, H., Ren, Y., Shang, P., Zhu, Y., Liang, Y., ... & Xu, D. (2019). The deep learning‐based recommender system "Pubmender" for choosing a biomedical publication venue: Development and validation study. *Journal of medical Internet research*, 21(5), e12957.

Goharian, N., El-Ghazawi, T., & Grossman, D. (2001). Enterprise text processing: A sparse matrix approach. *In Proceedings International Conference on Information Technology: Coding and Computing,* 71-75.

Goldstein, L. J., Lay, D. C., & Schneider, D. I. (2006). Calculus and its applications. *Prentice Hall*.

Gündoğan, E., Kaya, M., & Daud, A. (2023). Deep learning for journal recommendation system of research papers. Scientometrics, 128(1), 461-481.

Ham, J., Choe, Y. J., Park, K., Choi, I., & Soh, H. (2020). KorNLI and korSTS: New benchmark datasets for korean natural language understanding. *arXiv*. https://doi.org/10.48550/arXiv.2004.03289.

Kang, N., Doornenbal, M. A., & Schijvenaars, R. J. (2015). Elsevier journal finder: recommending journals for your paper. *In Proceedings of the 9th ACM Conference on Recommender Systems*, 261-264.

Lee, S., Jang, H., Baik, Y., Park, S., & Shin, H. (2020). Kr-bert: A small-scale korean-specific language model. arXiv. https://doi.org/10.48550/arXiv.2008.03979.

Liang, D., Charlin, L., McInerney, J., & Blei, D. M. (2016). Modeling user exposure in recommendation. *In Proceedings of the 25th international conference on World Wide Web*, 951-961.

Liu, C., Wang, X., Liu, H., Zou, X., Cen, S., & Dai, G. (2022). Learning to recommend journals for submission based on embedding models. *Neurocomputing*, 508, 242-253.

Lombardo, G., Tomaiuolo, M., Mordonini, M., Codeluppi, G., & Poggi, A. (2022). Mobility in unsupervised word embeddings for knowledge extraction－the scholars' trajectories across research topics. *Future Internet*, 14(1), 25.

Luong, H., Huynh, T., Gauch, S., Do, L., & Hoang, K. (2012a). Publication venue recommendation using author network's publication history. *In Intelligent Information and Database Systems: 4th Asian Conference (ACIIDS 2012)*, 426-435.

Luong, H. P., Huynh, T., Gauch, S., & Hoang, K. (2012b). Exploiting Social Networks for Publication Venue Recommendations. *In KDIR 2012 - Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*, 239-245.

Orkphol, K., & Yang, W. (2019). Word sense disambiguation using cosine similarity collaborates with Word2vec and WordNet. Future Internet, 11(5), 114.

Park, S., Moon, J., Kim, S., Cho, W. I., Han, J., Park, J., ... & Cho, K. (2021). Klue: Korean language understanding evaluation. *arXiv*. https://doi.org/10.48550/arXiv.2105.09680.

Pradhan, T., Gupta, A., & Pal, S. (2020). Hasvrec: A modularized hierarchical attention-based scholarly venue recommender system. *Knowledge-Based Systems*, 204, 106181.

Pradhan, T., & Pal, S. (2020). CNAVER: A content and network-based academic venue recommender system. *Knowledge-Based Systems*, 189, 105092.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv*. https://doi.org/10.48550/arXiv.1908.10084.

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), e1249.

Xia, P., Zhang, L., & Li, F. (2015). Learning similarity with cosine similarity ensemble. *Information sciences*, 307, 39-52.

Yang, Z., & Davison, B. D. (2012). Venue recommendation: Submitting your paper with style. *In 2012 11th International Conference on Machine Learning and Applications*, 681-686

Yu, S., Liu, J., Yang, Z., Chen, Z., Jiang, H., Tolba, A., & Xia, F. (2018). PAVE: Personalized Academic Venue recommendation Exploiting co-publication networks. *Journal of Network and Computer Applications*, 104, 38-47.

Abstract

Content-based Korean journal recommendation system using Sentence BERT

Yongwoo Kim* · Daeyoung Kim** · Hyunhee Seo** · Young-Min Kim***

With the development of electronic journals and the emergence of various interdisciplinary studies, the selection of journals for publication has become a new challenge for researchers. Even if a paper is of high quality, it may face rejection due to a mismatch between the paper's topic and the scope of the journal. While research on assisting researchers in journal selection has been actively conducted in English, the same cannot be said for Korean journals. In this study, we propose a system that recommends Korean journals for submission. Firstly, we utilize SBERT (Sentence BERT) to embed abstracts of previously published papers at the document level, compare the similarity between new documents and published papers, and recommend journals accordingly. Next, the order of recommended journals is determined by considering the similarity of abstracts, keywords, and title. Subsequently, journals that are similar to the top recommended journal from previous stage are added by using a dictionary of words constructed for each journal, thereby enhancing recommendation diversity. The recommendation system, built using this approach, achieved a Top-10 accuracy level of 76.6%, and the validity of the recommendation results was confirmed through user feedback. Furthermore, it was found that each step of the proposed framework contributes to improving recommendation accuracy. This study provides a new approach to recommending academic journals in the Korean language, which has not been actively studied before, and it has also practical implications as the proposed framework can be easily applied to services.

Key Words : Deep learning, Document similarity, Recommendation system, Research papers, SBERT(Sentence Bidirectional Encoder Representations from Transformers)

Received : June 14, 2023 Revised : July 6, 2023 Accepted : July 17, 2023

Corresponding Author : Young-Min Kim

* Department of Technology Management, Graduate School of Technology & Innovation Management, Hanyang University
** DPLANEX Corp.
*** Corresponding Author: Young-Min Kim
Department of Technology Management, Graduate School of Technology & Innovation Management, Hanyang University
222 Wangsimni-ro, Seongdong-gu, Seoul, 02455, Korea
Tel: +82-2-2220-2537, E-mail: yngmnkim@hanyang.ac.kr

저 자 소 개

김용우
서강대학교에서 경제학, 심리학 학사를, 독학사로 컴퓨터과학 학위를 취득했다. 이후 독일 Leuphana Universität Lüneburg에서 Management & Data Science 석사학위를 취득하였으며, 한양대학교 기술경영학 박사과정을 수료하였다. 주요 연구분야는 기계학습, 데이터마이닝, 자연어처리이다. 현재 교보생명그룹 데이터분석 전문법인 DPLANEX에서 Data Scientist로 근무 중이다.

김대영
The George Washington University에서 International Affairs를 전공하였다. World Vision Korea에서 Data Analyst로 재직했고, 현재 교보생명그룹 데이터분석 전문법인 DPLANEX에서 Data Analyst로 근무 중이다. 주요 관심분야는 머신러닝, 딥러닝이다.

서현희
한양대학교 산업공학과에서 학사학위를 취득하고 현재 교보생명그룹 데이터분석 전문법인 DPLANEX에서 Data Analyst로 근무 중이다.

김영민
한양대학교 산업공학과에서 학사, 석사 학위를 취득한 후 프랑스 Universite Paris-VI 컴퓨터공학과에서 석사, 박사 학위를 취득했다. 현재 한양대학교 기술경영전문대학원 교수로 재직하고 있다. 연구분야는 기계학습, 확률 그래프 모델, 정보 추출이다.