다차원 메타데이터 공간을 활용한 학술 문헌 추천기법 연구*

A Study on the Method of Scholarly Paper Recommendation Using Multidimensional Metadata Space

감미아 (Miah Kam)**
이지연 (Jee Yeon Lee)***

초 록

본 연구는 '우수한 성능의 메타데이터 속성 유사도 기반의 학술 문헌추천시스템'을 제안하는 데에 목적을 두고 있다. 본 연구에서는 정보조직에서 다루는 메타데이터의 활용과 계량정보학에서 다루고 있는 동시인용, 저자-서지결합법, 동시출현 빈도, 코사인 유사도의 개념을 활용한 문헌정보학 기반의 학술 문헌 추천기법을 제안하고자 하였다. 실험을 위해 수집한 '불평등', '격차' 관련 총 9,643개의 논문 메타데이터를 정제하여 코사인 유사도를 활용한 저자, 키워드, 제목 속성 간의 상대적 좌표 수치를 도출하였고, 성능 좋은 가중치 조건 및 차원의 수를 선정하기 위해 실험을 수행하였다. 실험 결과를 제시하여 이용자의 평가를 거쳤으며, 이를 이용해 기준노드와 추천조합 특성 분석 및 컨조인트 분석, 결과 비교 분석을 수행하여 연구질문 중심의 논의를 수행하였다. 그 결과 전반적으로는 저자 관련 속성을 제한 조합 혹은 제목 관련 속성만 사용하는 경우 성능이 뛰어난 것으로 나타났다. 본 연구에서 제시한 기법을 활용하고 광범위한 표본의 확보를 이룬다면, 향후 정보서비스의 문헌 추천 분야뿐 아니라 사회의 다양한 분야에 대한 추천기법 성능 향상에 도움을 줄 수 있을 것이다.

ABSTRACT

The purpose of this study is to propose a scholarly paper recommendation system based on metadata attribute similarity with excellent performance. This study suggests a scholarly paper recommendation method that combines techniques from two sub-fields of Library and Information Science, namely metadata use in Information Organization and co-citation analysis, author bibliographic coupling, co-occurrence frequency, and cosine similarity in Bibliometrics. To conduct experiments, a total of 9,643 paper metadata related to "inequality" and "divide" were collected and refined to derive relative coordinate values between author, keyword, and title attributes using cosine similarity. The study then conducted experiments to select weight conditions and dimension numbers that resulted in a good performance. The results were presented and evaluated by users, and based on this, the study conducted discussions centered on the research questions through reference node and recommendation combination characteristic analysis, conjoint analysis, and results from comparative analysis. Overall, the study showed that the performance was excellent when author-related attributes were used alone or in combination with title-related attributes. If the technique proposed in this study is utilized and a wide range of samples are secured, it could help improve the performance of recommendation techniques not only in the field of literature recommendation in information services but also in various other fields in society.

키워드: 추천기법, 학술 문헌, 메타데이터, 다차원 메타데이터 공간, 코사인 유사도, 유클리드 거리
recommendation methods, scholarly paper, metadata, multidimensional metadata space, cosine similarity, euclidean distance

---

* 이 논문은 연세대학교 문헌정보학과 박사학위논문 축약본임.
이 논문은 2022년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2022S1A5C2A0309359721)
** 연세대학교 문헌정보학과 강사(makiyma@hanmail.net) (제1저자)
*** 연세대학교 문헌정보학과 교수(jlee01@yonsei.ac.kr) (교신저자)
■ 논문접수일자 : 2023년 2월 15일 ■ 최초심사일자 : 2023년 3월 7일 ■ 게재확정일자 : 2023년 3월 13일
■ 정보관리학회지, 40(1), 121-148, 2023. http://dx.doi.org/10.3743/KOSIM.2023.40.1.121

1. 서 론

1.1 연구의 배경

정보 이용자의 선호를 분석하여 개별화된 아이템이나 정보를 제공하는 추천시스템(Adomavicius & Tuzhilin, 2005)은 다량의 정보 중 이용자의 관심사에 근접한 정보를 선별하여 주는 기능을 한다(Liling, 2019). 이와 같은 추천시스템은 연구 수행에 도움이 될 최적의 논문을 최대한 신속하게 확보하고자 하는 학술 연구자에게 중요한 의미를 지닌다.

학술 문헌의 생산 및 공유가 빠른 속도로 진행되면서(Khan et al., 2017) 연구자가 그 속도를 따라잡지 못하는 정보 과부하 현상이 점차 생기게 되었다. 이에 연구자에게 최적의 정보를 필터링해서 제공해줄 수 있는 추천시스템 개발에 관심이 쏠리고 있다. 연구자들에게 양질의 정보를 제공하여 정보검색에 소요하는 시간을 절약할 수 있도록 돕는 학술 문헌추천시스템은 1998년 Giles et al.의 CiteSeer Project 연구에서부터 비롯되었으며(Beel et al., 2016), 다양한 영역에서 그 유용성이 확인되는 가운데 양적 확대와 질적 발전이 거듭되고 있다(Chen & Lee, 2018).

이처럼 빅데이터 양상의 학술 문헌에 추천시스템을 적용하려는 시도가 양적으로 확대되고 있기는 하나, 질적인 측면에서는 개선과 보완의 여지가 있다. 기존의 학술 문헌추천시스템들을 일별하면, 대개 동일한 키워드 중심으로 추천하거나, 해당 문헌을 이용한 다른 이용자들이 함께 참조했던 문헌을 부연해서 추천하는 형태를 띤다. 연구주제와 키워드 및 저자 등의 다양한 정보들을 종합적으로 참조하여 추천되는 방식까지는 아직 이르지 못한 것이다. 즉 메타데이터의 다양한 특성을 복합적으로 활용하여 추천하는 시스템은 아직 찾아보기 힘든 상황이다. 최근 들어 Amhad & Afzal(2020), Waheed et al.(2019), Alshareef(2019) 등과 같이 메타데이터를 활용한 학술 문헌추천시스템 개발이 시도되지만, 이 역시 추천 과정을 살펴보면 각 메타데이터 요소의 유사도를 계산한 후 합계를 내거나 평균을 도출하여 제시하는 정도를 크게 벗어나지 못하고 있다.

국내 학술 문헌 추천기법에 관한 연구는 찾아보기가 더욱 어렵다(여운동 외, 2010). 최근 국내의 추천시스템 관련 연구 동향을 살펴본 결과 유튜브, 영화 추천, 리뷰와 상품추천 등과 같이 이용자의 선호도(프로필)를 활용한 미디어/상품 관련 추천 및 협업 필터링 연구가 주를 이루고 있었다. 학술 문헌 관련 국내 최신 연구로는 박대우 외(2020)의 빅데이터 기반 도서추천시스템 구축 연구, 원재상(2020)의 문맥을 고려한 논문 추천시스템 연구, 임윤정 외(2021)의 학술 빅데이터 기반 분야별 전문가 추천시스템 연구가 존재하였다. 하지만 이들 논문은 이용자 성향을 파악하여 추천하는 연구이거나, 한 속성(제목 혹은 키워드)만을 유사도 계산에 활용하였거나, 내용이 아닌 논문 심사자에 대한 추천을 다루고 있어 논문의 다양한 속성을 조합하여 활용하고자 하는 본 연구의 취지와는 차이가 있었다.

이에 본 연구는 지금까지의 국내외 논문과는 차별화된 연구를 수행하고자, 저자, 키워드, 제목 등 논문의 여러 메타데이터 속성을 복합적으로 조합시킨 코사인 유사도 기반 거리 측정

과 좌표 개념의 적용을 시도하였다. 특히 도서관의 정보서비스 현장에서 이루어지는 장서 추천서비스의 과정을 기반으로 하는 추천시스템 개발을 시도하고자 하였다.

본 연구는 한 이용자가 자료의 추천을 의뢰해 왔을 때 참고 사서가 추천인이 되어 방대한 자료 목록 가운데에서 의뢰인의 필요에 부합하는 최적의 자료를 찾아 추천하게 되는 과정을 상정하는 데에서 출발하였다. 추천을 위한 첫 단계는 이용자가 자신의 관심 주제와 관련이 있는 문헌을 추천해주기를 의뢰해오는 데에서 시작한다. 이를 접수한 사서, 즉 추천인은 의뢰인이 흥미로워한 문헌에 관하여 구체적으로 질문함으로써 추천의 기준이 될 정보인 연구주제, 키워드, 저자 등을 포함하는 정보의 조합을 구축하게 된다. 이를 바탕으로 확보 가능한 자료들 가운데 기준정보와의 전반적인 유사성이 높은 추천자료 후보군을 정한 후, 필터링하는 과정을 거쳐 최종적으로 최적의 자료들을 추천하게 되는 것이다. 저자만 같다거나, 키워드만 같은 경우를 생각해내는 것이 아니라, 예를 들면 '동일 저자가 쓴 비슷한 주제의 자료'와 같이 속성들을 동시에 판단하여 이를 조합한 후 추천하는 것이다.

이와 같은 절차는 본 연구에서 제시하고자 하는 메타데이터를 활용한 추천시스템 활용의 흐름과 잘 들어맞는다. 학술 문헌의 메타데이터를 확보한 후, 검색의 기준이 되는 문헌의 속성이 시스템에 입력되면, 입력된 다양한 메타데이터 속성조합과 유사도가 높은 다른 문헌이 추천되는 등의 순서로 진행될 수 있다는 것이다. 이 방식은 메타데이터의 단일 속성의 유사성에 따라 추천하는 방식에서 벗어나, 메타데이터의 여러 속성을 다차원의 공간에서 하나로 묶어 다룬다는 점에서 타 연구물과의 차별성을 지닌다.

본 연구는 계량정보학적 관점과 메타데이터의 특성을 기반으로 한 문헌정보학 지식체계를 활용하여 '메타데이터 속성 유사도 기반의 학술 문헌추천시스템'을 제안하는 데에 목적을 두고 있다. 본 연구는 문헌정보학 기반 전공 지식 활용을 통해 학술 문헌추천시스템의 성능을 높일 가능성을 찾고, 여러 학술 분야 및 시스템 환경에서 사용할 수 있는 발판을 마련하여, 문헌정보학 연구의 지평을 넓히는 데에 기여 하고자 하였다.

1.2 연구질문

본 연구의 연구질문은 다음과 같다.

- RQ1. 논문의 특성별로 효과적인 속성 및 추천조합이 달라지는가?
- RQ2. 다양한 특성을 가진 논문들을 두루 고려했을 때, 전반적으로 추천 성능이 뛰어난 추천조합은 무엇인가?
- RQ3. 기존의 추천시스템과 비교하였을 때 정보조직 및 계량정보학적 지식을 활용한 문헌정보학 기반 학술 문헌 추천기법의 효과적인 점과 한계는 무엇인가?

1.3 용어 정의

1) 속성조합(Attribute Combination; AC): 본 연구에서의 속성(Attribute)은 '저자', '제목', '키워드'이며, 속성 간의 조합을 속성조합이라 한다(〈표 1〉 참조). 속성조합

증가 시 차원도 증가한다. 예시로는 '키워드-동시출현'을 들 수 있는데, 한 논문 내에 동시에 출현한 키워드 간에는 유사성이 높음을 기본 개념으로 하여, 키워드 간 유사도를 도출해내었다. 두 논문의 키워드를 2개씩 비교하게 되면, 이 '키워드-동시출현' 속성조합은 2*2의 총 4개 축을 가지게 되며, 이로써 도출된 4개의 코사인 유사도는 향후 거리 측정에 사용된다.

2) 추천조합(Recommendation Combination; RC): 각 속성 및 속성조합들이 모여 이룬 조합으로, 추천에 사용될 방법들을 이른다. 예시로는 '저자-제목' 등과 같은 속성조합들이 모여 '저자-제목 & 저자-공저 & 저자-논문간 & 키워드(3-3) & 제목-논문간'으로 추천조합을 이룰 수 있다. 추천조합으로는 RC01-1~RC09의 총 50개 9개 군이 있으며(〈표 2〉 참조), 연구 전반에 걸쳐 이들 간의 성능을 비교하여 적합한 조합을 제시한다.

3) 기준노드: 현재 이용자가 보고 있는, 즉 추천의 기준이 되는 학술 문헌이며, 이를 기반으로 다른 노드와의 상대적인 거리를 계산하여 추천할 노드를 결정한다. 본 연구에서는 총 12개의 기준노드를 설정하고 실험을 수행하였다.

4) 추천노드: 기준노드의 가장 가까운 거리에 있는 노드로, 실제 추천이 되는 논문들이다. 본 논문에서는 top 5로 추천노드를 구성하여 결과를 제시하였으며, 필요에 따라 top 10의 추천노드도 분석하였다.

5) 후보노드: 추천노드로 제시될 가능성이 있는 표본 내 모든 논문을 의미하며, 실험 시 기준노드와의 거릿값을 산출하여 최종 추천노드가 선정된다.

6) 세타 거릿값: 코사인 유사도의 각도를 거릿값의 의미로 변환시킨 것으로, 속성을 나타내는 하나의 축 상에 노드 간 상대적인 거리를 나타내기 위해 사용한다. 이 세타 기반 거릿값들이 유클리드 공간 내에 여러 축을 이뤄 다차원으로 모이면, 이들을 활용하여 유클리드 거리를 계산한 후 최종 추천 논문을 도출해내게 된다.

7) 다차원 메타데이터 공간(Multidimensional Metadata Space): 메타데이터의 속성들(저자, 키워드, 제목)을 조합하여 각 '속성조합'으로 만든 후, 이를 각각 하나의 축으로 부여하였다. 이때 다양한 추천조합으로 인해 속성조합, 즉 축이 최대 22개까지 생성이 되므로 다차원을 이룰 수 있다. 각 축 간에는 직각을 이루고 있어 유클리드 공간에 기반한다. 이러한 다차원의 공간 내 위치한 하나의 노드는 다른 노드와의 유클리드 거리를 가지게 되고, 이 거리를 산출하여 가장 가까운 상위 노드를 추천하는 방식이 된다. 이 공간의 축은 문헌의 특성이나 이용자들이 찾고자 하는 방향성에 따라 그 수가 변경될 수 있다.

한편 추천의 기준이 되는 기준노드의 경우 이 공간상에서 항상 원점을 지닌다. 코사인 유사도의 경우 속성 간 짝지어진 유사도만을 보여주기 때문에, 개념상의 다차원 공간 위에 절대적인 값으로 두기가 어려웠다. 이런 문제를 해소하기 위해 기준노드와 대상 노드 속성의 상대적인 거리(세타값)를 활용하여 기준노드

와의 거리를 산출하는 방식을 고안해내었다. 즉 다차원 공간상에 각 노드의 좌표가 어딘가에 존재한다고 가정하고서, 미지의 절대적 좌표 대신 노드 간의 상대적인 좌표(코사인 유사도 기반 각도, 즉 거릿값)를 사용하여 기준노드를 원점으로 설정한 후, 이와 모든 후보노드 간의 거리를 구하고자 하였다. 이를 본 연구에서는 '메타데이터'를 기반으로 한 '다차원 공간'에서의 거리 측정이라는 개념으로 설정하여, 추천이 이루어지는 공간을 '다차원 메타데이터 공간'이라 명명하였다.

2. 이론적 배경

학술데이터는 수백만 명의 저자, 인용, 그림, 표뿐만 아니라 학술 네트워크, 디지털 도서관과 같은 대규모 프로젝트 관련 데이터를 포함한다(Xia et al., 2017). 학술데이터는 세부적으로 학술지 기사, 회의 절차, 학위논문, 서적, 특허, 프레젠테이션 슬라이드 및 실험데이터와 같은 학술 활동 결과 나타난 데이터를 뜻하는데(Williams et al., 2014), 이 중 본 연구는 학술지 기사를 뜻하는 '학술 문헌'에 집중하여 수행되었다.

추천시스템(Recommendation System)은 이용자의 선호를 분석하여 개별화된 아이템이나 정보를 이용자에게 제공하는 시스템으로(Adomavicius & Tuzhilin, 2005), 다양한 연구들이 있지만 본 연구와 비슷한 방법을 쓴 연구 몇 가지로 추려 제시하였다. 특히 학술 문헌 중 메타데이터(저자, 키워드, 제목)와 코사인 유사도를 중점적으로 활용한 연구를 위주로 제시하였으며, 이들과의 차이점을 분석해 본 연구만의 특성과 의의를 찾고자 하였다.

Waheed et al.(2019)의 연구에서는 저자에 기반하여 논문을 추천하기 위해 인용 네트워크와 저자 랭킹을 사용한 혼합 접근법을 활용하였다. 인용 정도는 '많이 이용된'의 의미로 접근하여 활용되었다. 이 논문에서는 세 가지 기본 단계를 가지는데, 방향성 있는 다층 인용 네트워크(directional multilevel citation networks)를 생성하는 최초의 단계, 각 논문의 후보 점수를 계산함으로써 후보 논문들을 선택하는 두 번째 단계, 그리고 최종 추천을 위하여 후보 논문의 순위를 평균하는 마지막 단계로 구성되어 있다. 이 연구에서는 공저자 분석을 통해 네트워크를 만든 후 다시 중심성이 높은 저자를 선택하고, 그 저자가 쓴 논문을 이용자에게 추천하는 형태로 이루어져, 인기 있는 논문을 추천하는 데 의의를 두고 있었다.

Google Scholar는 본 논문의 비교 분석에서도 활용되기에, 이를 검토한 Beel & Gipp(2009)의 연구를 제시한다. Google Scholar는 주로 인용과 공저에 기반한 통계모델을 이용하고 있다. 여기에는 Google의 PageRank가 이용되었는데, 이는 일종의 투표행위처럼 다른 학술논문들로부터 받은 인용횟수에 기반하여 권위를 측정하고 순위를 매기는 것이다. 연결되어 들어오는 페이지가 많을수록, 높은 순위의 노드가 들어오는 링크가 많을수록 높아지는 방식이다. 이 방식은 최근에 출판된 논문이 관심 논문으로 선택되는 경우 양질의 논문들을 추천하는 데 실패할 수 있다.

위의 연구들과 같이 논문의 인용 네트워크 자체를 활용하면 인기 있는 논문과 저자 중 상위

저자를 찾아 제시하는 방식 등을 택하기 때문에, 새로 나온 아이템이나 인기 없는 논문에 대한 추천이 어렵다는 한계를 가진다. 그리고 연구물의 영향을 평가하기 위해 영향력 지수, 인용횟수, 동시인용 등과 같은 다양한 변수를 활용하는 방식은 단지 하나의 수준(저자)만을 고려하는 것으로, 또 다른 요인들의 효과를 반영하지 않는다고 볼 수 있어 한계가 있다(Alshareef, 2019). 이처럼 인용 네트워크 한 가지를 주로 활용하는 사례도 있고, 이러한 cold-start의 한계를 극복하고자 다음과 같이 메타데이터를 활용하여 추천을 시도하는 방법들도 나타난다.

메타데이터는 처음 생성 시 시스템상에서 전문을 파악하고 이에서 자동으로 가져오는 방식을 택하거나 키워드의 경우 저자가 직접 작성, 혹은 메타데이터 관리자가 수동으로 등록하는 경우가 대부분이다. 그래서 생성하는 도중에 오류가 생길 수 있고 생성 방식 또한 일관성이 떨어질 수밖에 없다(Deldjoo et al., 2019). 그럼에도 메타데이터를 활용하는 이유는 앞서 언급한 인용기반 시스템의 cold-start 문제에 대한 해결 가능성과 전문 대체재로서의 가능성이 발견되고 있기 때문이다. Nazir, Asif, & Ahmad(2020)는 전문으로부터 제목 및 초록을 추출하여 메타데이터의 핵심 용어와 완성된 콘텐츠에 코사인 유사도를 적용했다. 그 결과, 0.68의 유사도 점수를 얻을 수 있었고, 이로써 메타데이터는 전체 연구 작업의 상당 부분을 정확한 형태로 표현하고 있으며 연구 목적 및 조작에 활용될 수 있음을 확인하였다.

Alshareef(2019)는 박사학위논문으로 메타데이터를 활용한 추천기법을 구체적으로 제안하였는데, 여기에서는 참고문헌 정보에 대한 코사인 유사도를 기반으로 한 매트릭스를 적용했다. 이 논문은 논문, 저자, 논문 출판 방식(학술대회/학술지)의 인용 영향을 평가하기 위한 계량서지학적 지수들을 내용 유사도와 결합하여 두 개체 사이의 의미론적 유사도를 계산한 학술추천모델을 제안하였다. 이 논문은 논문, 저자, 논문 출판 방식이라는 세 가지의 요소와 계량서지학적 지수를 활용하였다는 점에서 본 연구와 접근 방식이 비슷하다. 그러나 추천의 주된 목적이 논문뿐만 아니라 저자 및 논문 출판 방식도 찾기 위함이라는 점과 각 논문의 영향력을 전반적으로 중요시하였다는 점에서 본 연구와 차이가 있었다.

Ahmad & Afzal(2017)은 인용 정보를 기본으로 하여, 제목, 저자, 키워드 중 하나, 둘, 혹은 전체를 조합하여 유사도의 측정치를 더한 값으로 추천을 수행하는 방식을 채택했다. 즉 전통적인 동시 인용이 메타데이터 적합성과 결합될 때 더 나은 결과를 낼 것이라는 가설로 시작했는데, 이를 위해 메타데이터 필드의 유사도 값을 계산하고 이를 결합하여 추천 결과를 제시하였다. 다양한 조합으로 실험한 결과, 키워드까지 포함하는 것보다는 “동시 인용+제목+저자”에서 성능이 더 좋게 나타났다.

Ahmad & Afzal(2020)은 위의 연구를 추가로 발전시켰는데, 이 연구에서는 제목, 초록, 키워드 메타데이터를 동시 인용에 통합하였다. 여기에서의 동시 인용 및 제목-초록의 결합도는 코사인 유사도 점수를 평균 내어 도출하는 방식을 택하였다. 이 논문에서의 기법을 활용한 결과 제목과 초록만을 결합함으로써 0.81의 NDCG(Normalized Discounted Cumulative Gain) 점수를 달성할 수 있었기에, 이 분야에

서는 제목과 초록의 결합을 통한 추천이 잘 작동한다는 사실을 알아내었다.

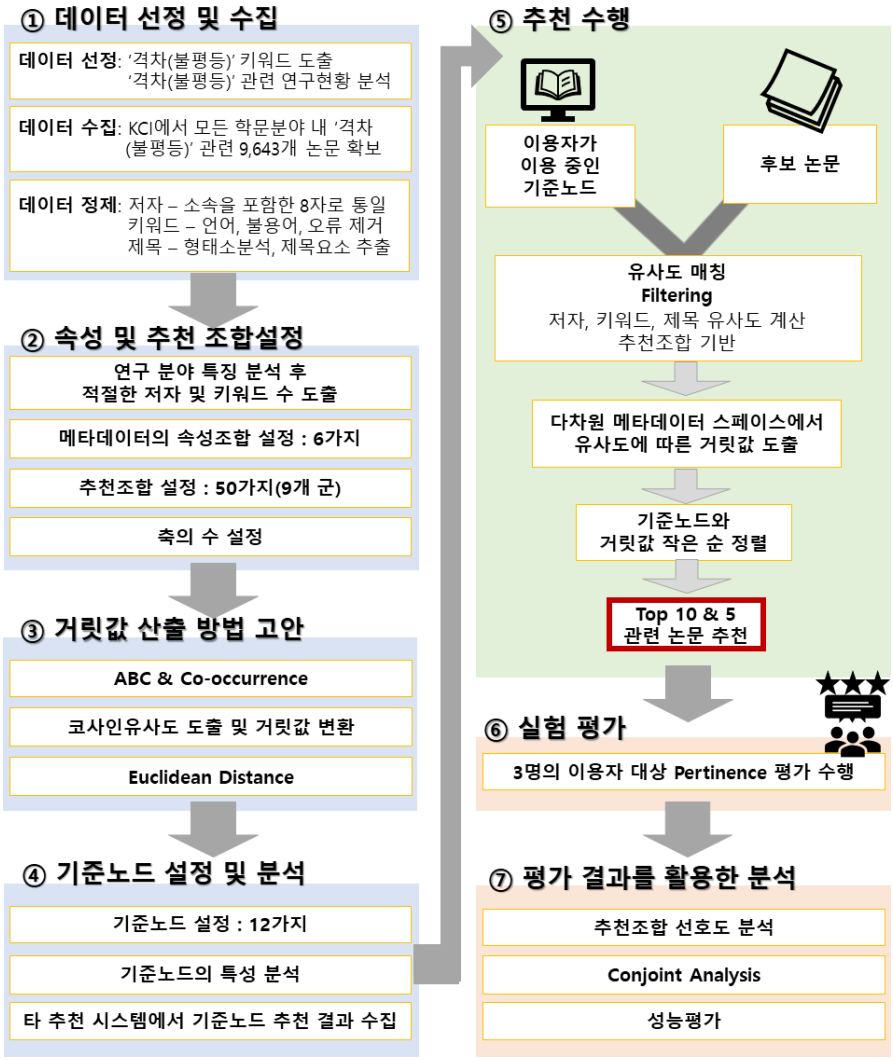
본 장을 통해 본 연구과 밀접히 관련된 국내외의 최신 연구들을 살펴보았다. 주목할만한 점은 특히 근래에 들어 전반적으로 인용 정보를 기본적으로 활용하고 있으며 그것이 매우 중요한 역할을 하고 있다는 것이었다. 또한 본 연구와 비슷하게 코사인 유사도를 활용한다거나, 메타데이터의 정보 등을 활용하여 추천시스템을 새롭게 제안하고 있었다. 특히 Ahmad & Afzal(2020)의 연구는 저자, 키워드 메타데이터 속성을 사용한다는 점과 이 메타데이터 속성들을 조합한다는 점에서 본 연구와 접근 방식이 비슷하였다. 그러나 들여다보면, 대부분 코사인 유사도 점수를 더하여 제시하는 정도에서 그치고 있었다. 본 연구에서는 평균 정보와 분포정보를 같이 이용하여, 그보다 나은 추천 결과를 보이고자 통합과 거리 개념을 활용하였다. 속성 간 더하기를 통한 단일 차원 내에서의 비교보다는, 다차원 내에서 평균 및 분포 정보를 활용하게 되면 균형 잡힌 추천이 이루어진다고 보았다. 이처럼 종합적인 접근으로 추천을 시도했다는 점에서 최신 연구들과의 차이가 있었다. 또한 선행연구의 경우 저자 관련 데이터는 사용하지 않고, 초록과 제목의 결합을 시도하였다는 점에서 본 논문과 차이가 있다. 더불어 논문 특성별로의 세부적인 추천 방식에 대한 논의는 없었는데, 다양한 논문의 특이사항에 대해 다루는 것이 필요하다고 판단하였다.

이에 본 연구에서는 메타데이터의 속성조합에 대한 장점을 파악하고, 논문 특성별로는 세부적인 분석, 전체 논문의 차원에서는 종합적으로 분석하여, 논문 특성에 따른 적합한 추천 방식을 아우르는 새로운 기법을 제시하고자 한다.

3. 연구방법

다음 〈그림 1〉은 본 연구의 절차로, 각 그림 옆의 숫자를 따라 방법론을 설명하고 있다. 〈그림 1〉에서의 ①~④의 경우 추천을 수행하기 전 준비하는 단계이며, ⑤는 실제 추천이 수행된 절차, ⑥은 그에 따른 이용자들의 평가, ⑦은 이용자들의 평가 결과를 활용한 제안된 추천기법에 대한 분석이다. 전반적 연구 수행의 과정을 요약하면 다음과 같다.

우선 주요 주제 선정을 통해 실험을 위한 총 9,643개 논문을 확보하였고, 논문의 메타데이터에서 나타나는 제목에 대한 형태소 분석과 불용어 제거, 키워드 및 저자 관련 메타데이터를 수정하여 데이터 전처리를 하였다(①). 그런 후 데이터의 특징 및 메타데이터의 속성을 분석하고, 각 메타데이터 속성조합 설정 및 추천조합을 설정한 후(②) 거릿값 산출에 대한 방법을 고안하였다. 이를 위해 코사인 유사도 값을 통해 노드 간의 상대적인 거릿값을 계산하였고(③), 12개의 표본 논문(기준노드)을 설정하였다. 이 12개 논문에 대해 타 시스템의 추천 결과를 수집하여 추천을 위한 전반적인 준비를 끝낸 후(④), 실제 추천을 수행해보았다(⑤). 추천 수행 이후 3명의 이용자로부터 결과에 대한 전반적인 평가를 받았다(⑥). 그런 후, 이 평가 결과로 추천조합별 평가 점수 및 순위를 제시했고, 추천조합에 대한 종합적인



〈그림 1〉 연구 절차

순위를 이용자별로 매겨보았다. 이 순위를 활용하여 이용자들이 선호 조합을 도출해내는 컨조인트 분석을 수행하였고, 전체적으로 적합한 조합을 분석하였다. 그런 후 타 시스템과의 결과 비교를 수행하여 성능평가를 했고, 학술 문헌 추천 실험에 대한 종합적인 분석 및 연구질문 기반의 논의를 하였다(⑦).

3.1 데이터 수집 및 정제

격차(불평등)는 다양한 분야에 걸쳐 오랜 시간 논의가 되어오고 연구가 되어온 분야이며, 같은 용어라도 학문 분야에 따라, 혹은 해당 용어 앞에 붙는 용어(정보, 소득, 건강, 문화 등)에 따라 다른 쓰임이 된다. 이에 격차(불평등)

는 다른 키워드들이 쉽게 가질 수 없는 특성을 띠는, 사회과학 내 존재하는 특별한 용어라 보았다. 본 연구에서는 사회과학일반, 각 학문분야, 그리고 문헌정보학의 세 측면에서 모두 의미 있게 다루고 있는 것으로 보이는 '격차(불평등)'를 주된 키워드로 다루고자 하였다. 격차(불평등) 관련 연구의 건수나 비율로 보았을 때 경제학 및 사회복지학 분야의 영향이 크다고 보아, 본 연구의 추천 결과에 대한 이용자 평가 시 경제학과 사회복지학 전공의 박사급 연구자들로 구성하고자 하였다.

KCI 웹페이지(한국연구재단, 2020)를 활용하여 학술 문헌의 메타데이터를 확보하였다. 데이터의 수집 시점은 2020년 4월 19일이며, KCI 논문검색에서 '불평등 or 격차'로 상세검색하여, '자료반출'을 통해 사회과학 분야와 문헌정보학을 포함한 '모든 학문 분야'의 총 9,643개 논문에 대한 메타데이터를 확보하였다. 메타데이터의 속성으로는 전체를 확보했는데, 논문제목과 저자명, 저자소속기관, 발행기관명, 학술지명, 키워드가 포함되었다. 이 중 기재가 온전히 되어있었던 저자 메타데이터의 수는 총 9,643개이고 총 저자의 수는 10,855명이다. 키워드 메타데이터 수는 9,577개에 키워드 수는 총 28,770개이다. 제목 메타데이터는 9,306개이며 제목 요소 수는 9,533개이다. 논문들은 모두 한국에서 출판된 논문이기에 언어는 한글을 기본으로 하였다.

기존의 메타데이터 상에 저자의 이름과 소속 외에도 저자의 고유 코드가 제공된다면, 그 고유코드를 활용하여 저자 처리를 할 수 있을 것이다. 저자의 고유 코드로 분석할 시 저자 이름 및 소속이 같을 때 나타나는 동명이인 문제는 상당히 감소할 것으로 보인다. 하지만 확보한 메타데이터에는 저자의 소속기관 정보만 나와 있어, 동명이인 문제를 해결하기 위해 저자명뿐만 아니라 소속 정보도 함께 사용하였다. 아울러 저자 속성 내에서의 통일성을 확보하고자 총 8자리까지 처리하였다. 예를 들면, '홍길동(서울대학교)'의 형태로 제시된 경우, '홍길동(서울대학)'으로, '가나다(연세대학교)'는 '가나다(연세대학)'으로 처리하여 같은 소속 동명인에 대해 동일인으로 처리하도록 하였고, 동명이지만 소속이 다를 시 다른 저자로 처리하였다.

키워드는 KCI에서 메타데이터로 구축한 학술논문의 저자키워드를 사용하였다. 본 논문에서는, 대개 동일한 저자키워드를 공유하는 경우 대체로 유사한 개념 영역에 속하게 된다(고영만, 송민선, 이승준, 2015)는 개념을 활용하고자 하였다. 논문의 키워드 처리를 위해서는, 전체적으로 훑으며 오류를 최소화하려 하였다. 우선 단순 빈도로 가장 많이 등장한 키워드와, '불평등', '평등', '격차' 관련된 키워드는 분석에 영향을 줄 수 있으므로 삭제하고, 불용어 및 각종 기호를 제거하였다. 영어와 한국어가 혼재해 있는 경우 영어를 제거하고 한국어 키워드만 남겨놓았으며, 한국어와 영어의 뜻이 다른데 함께 나타난 경우는 영어도 남겨두었다. 영어뿐만 아니라 한자, 프랑스어, 일본어, 독일어 등 한글 키워드와 중복되는 의미로 존재하는 외국어 키워드의 경우 모두 제거되었다. 키워드 추출 후 개별 키워드 전체에 대해 가나다순 및 출현빈도순으로 정렬하여, 동형이의어의 문제가 없는지 연구를 진행하는 동안 살펴보았다. 관련 전공 두 박사과정의 연구원이 이를 비교하여 확인하였으며, 동일한 의미를 지닌 용어

의 경우 두 명이 공통으로 인정하는 키워드를 선정하여 이로 통일시키는 작업을 수행하였다.

메타데이터 상의 오류가 상당히 있었는데, 키워드 메타데이터 부분에 초록이 들어가 있거나, 목차가 들어가 있는 경우, 혹은 문헌분류기호나 권제, 감사의 글이 들어가 있는 경우까지 오류의 유형은 다양했다. 이와 같은 오류는 결과에 영향을 미칠 수 있어 모두 제거하였다.

제목 요소 처리를 위해 제목으로부터 키워드 형식의 요소를 추출하였고, 형태소 분석을 통해 고유명사와 일반명사를 사용하였다. 추가로 '어근 뒤에 명사파생접미사가 따라오는 경우'나, '체언접두사에 일반명사가 따라오는 경우' 등 형태소끼리 조합하였을 때 제목 요소로 의미가 있는 용어들은 결합하여 활용하였다.

3.2 속성조합 설정

본 절에서는 조합 설정 시 적합한 저자 및 키워드 고정값을 결정하고자 저자 수와 키워드 수를 살펴보았다. 전체 확보한 논문 9,643개 중 단일저자인 경우가 다수인 상황(약 62%)이며, 평균적으로는 한 논문 당 약 1.55명의 저자가 있다는 사실을 알 수 있었다. 이를 통해 저자 정보를 추천에 적용할 시 첫 번째 저자까지, 혹은 더 넓게는 두 번째 저자까지 고려하는 것이 효과적임을 예상할 수 있었다. 키워드 수를 보면 보통 한 논문 당 2~4개 수준으로 구성되어 있었기에 추천에 활용할 적절한 키워드의 수는 3개로 선정하여 실험에 적용하였다.

최종적으로 확보한 학술 문헌 메타데이터 속성 중 본 논문에서는 '저자', '키워드', '제목'의 메타데이터를 활용하기로 하였고, 이 속성들은 속성조합의 개념으로 설정하여 각 속성조합의 처리 방식에 따라 각기 다른 저자 유사도, 키워드 유사도, 제목 유사도를 가지도록 하였다. 다음의 〈표 1〉과 같이 '속성1: 저자', '속성2: 키워드', '속성3: 제목'을 메타데이터 세 가지 속성으로 두었으며, 축을 이루는 속성조합은 셀에 나타나 있다.

다음은 〈표 1〉에서 제시하는 각 속성조합에 대한 설명이다.

- '속성조합1: 저자-제목': 제목 요소의 유사성에 따른 저자 간 유사도로, 저자들 간 동일한 제목 요소를 쓸수록 유사도가 증가하는 형식이다.
- '속성조합2: 저자-공저': 공저의 정도에 따른 저자 간 유사도로, 두 저자 간에 공동저자로 등장하는 정도를 분석하여 유사도를 계산하였다.

〈표 1〉 메타데이터 속성에 따른 속성조합

	속성조합		
속성1: 저자	속성조합1 저자-제목	속성조합2 저자-공저	속성조합3 저자-논문간
속성2: 키워드	속성조합4 키워드-동시출현	속성조합5 키워드-논문간	-
속성3: 제목	속성조합6 제목-논문간	-	-

- ‘속성조합3: 저자-논문간’: 논문 간에 동일한 저자가 존재할수록 논문 간 유사도가 증가하는 형식이다.
- ‘속성조합4: 키워드-동시출현’: 한 논문 안에서 동시에 출현한 정도에 따라 키워드 간의 유사도를 계산하여 제시하였다.
- ‘속성조합5: 키워드-논문간’: ‘저자-논문간’과 동일하게, 논문 간 동일한 키워드가 다수 존재할수록 논문 간 유사도가 증가하게 된다.
- ‘속성조합6: 제목-논문간’: 논문 간 동일한 제목 요소의 수에 따라 계산된 논문 간 유사도이다.

‘속성조합1: 저자-제목’, ‘속성조합3: 저자-논문간’과 ‘속성조합5: 키워드-논문간’, ‘속성조합6: 제목-논문간’의 경우 Latent Semantic Analysis(LSA) 및 저자-서지결합법와 관련하여 근래 추천시스템 및 정보검색 분야 등에서 활용에 대해 논의되고 있는 부분과 연관되어 있다.

기존 연구에서는 제목에 포함된 단어들을 통해 논문 간의 latent semantic 관계를 분석하거나(Alshareef, 2019), 공통 용어가 포함된 논문 간에는 공통적인 연구주제를 가지고 있다고 암시한다고 보는 시각에서(Morris & Yen, 2004) 저자 수준으로 키워드 동시 발생 관계를 확장하고 저자 관계를 설정하는 저자-키워드 커플링 분석을 도입하는 연구(Yang et al., 2016)도 있었다.

이러한 기존 연구에서의 제목-논문의 유사도 구하는 방식과 관련하여 본 연구에서는 저자, 키워드, 제목 구성을 통해 논문 간 유사도를 구하는 방법을 속성조합으로 제안하였고, 기존의 저자-키워드 커플링 분석과 관련하여 ‘속성조합1: 저자-제목’을 제안하였다.

코사인 유사도 도출을 위해 초반에 처리하는 매트릭스는 속성조합별로 다음과 같이 설명할 수 있다. 우선 ‘속성조합1: 저자-제목’의 경우는 초반에 셀 내에 저자가 활용한 제목 요소 빈도를 나타낸 저자×제목 비대칭 행렬을 사용하여 저자-저자 대칭 유사도행렬을 도출한다.

‘속성조합2: 저자-공저’ 및 ‘속성조합4: 키워드-동시출현’은 하나의 문헌 내에서 동시에 출현하는 정도에 따라 대칭행렬을 가진다. 이는 다시 저자-저자 유사도 행렬 및 키워드-키워드 유사도행렬을 만들게 된다.

‘속성조합3’, ‘속성조합5’, ‘속성조합6’인 저자, 키워드, 제목의 ‘논문 간 유사도’는 논문×제목, 논문×저자, 논문×키워드 비대칭 행렬을 사용한다. 이를 이용하여 논문-논문 유사도행렬로 만들어 코사인 유사도를 각 셀 내에 기재하는 방식으로 이루어졌다.

‘속성조합1 & 속성조합2’의 경우 저자 간 유사도를 나타내는데, 이때 ‘속성조합1: 저자-제목’에서 두 명의 저자까지 고려하여 유사도를 도출하면 총 4차원의 공간이 만들어진다. 기준노드와 후보노드 간의 저자 유사도를 고려하는 방법이기에, 두 명까지 고려한다는 것은 기준노드의 첫 번째 저자와 후보노드의 첫 번째 저자, 기준노드의 두 번째 저자와 후보노드의 첫 번째 저자, 기준노드의 첫 번째 저자와 후보노드의 두 번째 저자, 기준노드의 두 번째 저자와 후보노드의 두 번째 저자 조합을 본다는 의미가 되고, 이는 총 2x2의 유사도를 보는 경우이므로 총 4개 축이 이루어진다(이때 표현은 ‘저

자-제목(2-2)'로 함). 여기에 더불어 '속성조합2: 저자-공저' 역시 두 명의 저자까지 고려하여 '속성조합1'과 결합하면 같은 방식으로 인해 총 8차원의 공간이 만들어진다. 이를 그림으로 표현하면 다음 〈그림 2〉와 같다.

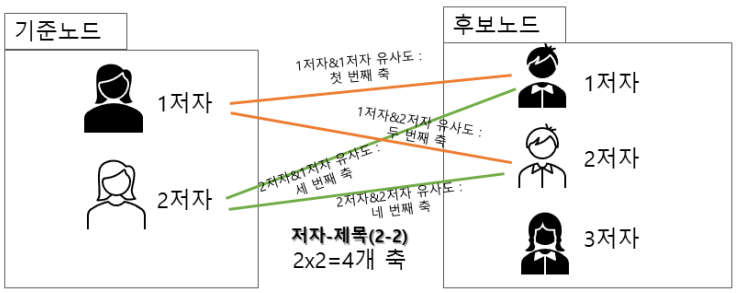
'속성조합4: 키워드-동시출현'도 마찬가지로 이루어진다. 기준노드에서 3개의 키워드와 후보노드에서 2개의 키워드를 조합하고자 하는 경우, 이때는 '키워드-동시출현(3-2)'와 같이 표현하며, 총 3x2의 6개 축이 구성되는 형식이다.

'저자-제목(2-2)' 속성조합의 예시처럼 2x2의 저자 방법만을 활용할 때, 기준노드에 '두 명의 저자'가 존재하지만 어떤 후보노드에는 '단일저자'만 존재할 경우, 그 후보노드는 두 번째 저자에 대한 유사도 값을 가지지 못하여 상대적으로 두 번째 저자 이상의 다른 후보노드에 비해 불리한 위치에 있게 된다. 이와 같은 단점을 해결하고자, '저자-제목(2-1)' 속성조합을 제안하였다. 이는 후보노드 저자 수에 상관없이, 후보노드 첫 번째 저자와의 유사도만 확인함으로써 단일저자를 가진 노드의 불리함(penalty)을 없애주는 방식이라 볼 수 있다. 또한 기준노드가 단일저자를 가질 때에도 '저자-제목(2-1)'의 형태라면 기준노드의 단일저자와 다른 후보노드의 첫 번째 저자와의 유사도 값만 도출되는 형태('저자-제목(1-1)'과 동일)가 되고 기준노드의 두 번째 저자 관련된 축은 없어진다. 초반에 2-1로 설정해놓으면 기준노드의 저자 수에 따라 1x1축이 나타나기도 하고 2x1축이 나타나기도 할 것이기에, 기준노드의 특성에 따라 더 많은 축을 활용할 수 있는 여지가 생긴다는 점에서 '저자-제목(1-1)'보다는 추천 성능이 좋을 것으로 예상하였다.

3.3 추천조합 설정

속성조합을 조합하여 추천시스템에 활용할 수 있는 형태로 구체화하면 추천조합이 된다. 이는 실제 추천 결과에 가장 큰 영향을 주는 조합이라 세부적인 조합 설정이 중요하다.

추천을 수행하기 위해서는 〈표 1〉에서 제시한 조합의 여섯 가지를 모두 포함하여 유사도를 종합할 수도 있고, 한두 가지만 포함할 수도 있다. 예를 들면 어떤 추천조합의 경우 '속성조합1 & 속성조합2 & 속성조합3'만을 포함하여 저자의 영향만을 고려할 수도 있고, 어떤 추천조합의 경우 '속성조합4 & 속성조합5'만을 가져 키워드의 영향만을 고려할 수도 있다.



〈그림 2〉 '저자-제목(2-2)' 속성조합의 개념 및 축 도출 방법

본 연구에서는 다양한 추천조합을 실험하고 각 논문 특성에 맞는 적합한 조합을 발견하고자, 다음 〈표 2〉의 50가지의 추천조합을 제시하여 분석했다.

RC(Recommendation Combination; 추천조합)의 경우 속성조합의 특성별로 순서화되어 있다. 여기에서 RC01군의 경우 저자, 키워드, 제목을 각각 한 축씩만 부여하고 각 속성 내 유사도의 평균을 내어 3차원으로 표현하는 방식인데, 이는 속성조합에 대한 처리의 변형에 해당한다. 한 예로, 표에서의 RC01-01은 저자, 키워드, 제목의 총 3차원의 공간을 활용하여 추천을 수행하는데, 저자축의 값으로는 Average('저자-제목(2-1)' & '저자-공저(2-1)' & '저자-논문간'), 키워드축의 값으로는 Average('키워드-동시출현(3-3)' & '키워드-논문간'), 제목축의 값으로는 Average('제목-논문간')를 가지게 된다.

표에서 비어있는 셀은 해당 속성조합을 활용하지 않는다는 뜻으로, RC09의 경우 저자 셀들과 키워드 셀들이 비어있고 '제목-논문간'에만 O 표시가 되어있으므로 '제목 속성'만 활용한 1차원의 추천조합이라는 뜻이 된다.

이는 전체 축(차원)의 수를 결정하는 방식을 결정짓는 것이라 볼 수 있다. 1명의 저자만 활용하여 추천 결과를 도출할 것인지, 3개 키워드까지 적용할 것인지에 따라 축의 수가 증감한다. 즉 두 번째 저자까지 활용하면 2*2=4개의

〈표 2〉 50가지 추천조합 설명

RC군	RC ID	추천조합 구성					
		저자-제목	저자-공저	저자-논문간	키워드-동시출현	키워드-논문간	제목-논문간
RC01군	RC01-01	(평균) 저자-제목(2-1) & 저자-공저 (2-1) & 저자-논문간			(평균) 키워드-동시출현(3-3) & 키워드-논문간		O
	RC01-02	(평균) 저자-제목(2-2) & 저자-논문간			(평균) 키워드-동시출현(3-3) & 키워드-논문간		O
	RC01-03	(평균) 저자-제목(1-1) & 저자-논문간			(평균) 키워드-동시출현(3-3) & 키워드-논문간		O
	RC01-04	(평균) 저자-제목(1-1) & 저자-논문간			(평균) 키워드-동시출현(3-2) & 키워드-논문간		O
	RC01-05	(평균) 저자-제목(1-1) & 저자-공저 (1-1) & 저자-논문간			(평균) 키워드-동시출현(3-3) & 키워드-논문간		O
RC02군	RC02-01			O		O	O
	RC02-02			O	(3-3)	O	O
	RC02-03	(2-1)	(2-1)	O		O	O
	RC02-04	(1-1)	(1-1)	O		O	O
	RC02-05	(1-1)		O		O	O

RC군	RC ID	추천조합 구성					
		저자-제목	저자-공저	저자-논문간	키워드-동시출현	키워드-논문간	제목-논문간
RC03군	RC03-01	(5-1)	(5-1)	O	(3-3)	O	O
	RC03-02	(3-1)	(3-1)	O	(3-3)	O	O
	RC03-03	(2-2)	(2-2)	O	(3-3)	O	O
	RC03-04	(2-1)	(2-1)	O	(3-3)	O	O
	RC03-05	(1-2)	(1-2)	O	(3-3)	O	O
	RC03-06	(1-1)	(1-1)	O	(3-3)	O	O
	RC03-07	(1-1)		O	(3-3)	O	O
	RC03-08	(1-1)		O	(3-2)	O	O
	RC03-09			O	(3-2)		O
	RC03-10		(1-1)			O	O
	RC03-11	(5-1)	(5-1)	O	(3-2)	O	O
	RC03-12	(3-1)	(3-1)	O	(3-2)	O	O
	RC03-13	(2-2)	(2-2)	O	(3-2)	O	O
	RC03-14	(2-1)	(2-1)	O	(3-2)	O	O
	RC03-15	(1-1)	(1-1)	O	(3-2)	O	O
RC04군	RC04-01	(2-2)	(2-2)	O	(3-3)	O	
	RC04-02	(2-1)	(2-1)	O	(3-3)	O	
	RC04-03	(1-1)	(1-1)	O	(3-3)	O	
	RC04-04	(2-2)	(2-2)	O	(3-2)	O	
	RC04-05	(2-1)	(2-1)	O	(3-2)	O	
	RC04-06	(1-1)	(1-1)	O	(3-2)	O	
	RC04-07	(1-1)		O	(3-3)	O	
	RC04-08	(1-1)		O	(3-2)	O	
	RC04-09	(1-1)		O		O	
	RC04-10	(1-1)	(1-1)		(3-2)		
RC05군	RC05-01	(2-2)	(2-2)	O			O
	RC05-02	(2-1)	(2-1)	O			O
	RC05-03	(1-1)	(1-1)	O			O
	RC05-04	(1-1)					O
RC06군	RC06-01	(2-2)	(2-2)	O			
	RC06-02	(2-1)	(2-1)	O			
	RC06-03	(1-1)	(1-1)	O			
	RC06-04		(1-1)	O			
RC07군	RC07-01				(3-3)	O	O
	RC07-02				(3-2)	O	O
RC08군	RC08-01				(3-3)	O	
	RC08-02				(3-3)		
	RC08-03				(3-2)	O	
	RC08-04				(3-2)		
RC09군	RC09						O

축이, 3개 키워드까지 활용하면 3*3=9개의 축이 생기는 것이다.

3.4 거릿값 도출을 통한 추천 수행

각 속성조합에 대한 코사인 유사도 값이 도출되면, 이를 통해 추천조합의 거릿값이 도출된다. 그러므로 추천조합이 어떻게 이루어지느냐에 따라 메타데이터 스페이스 상의 차원 수가 달라지고, 그에 따라 추천조합의 거릿값이 달라진다.

순서는 다음과 같다. 서로 다른 스케일을 가지는 출현빈도에 대해 두 속성 간 '① 코사인 유사도 도출'을 하여 0에서 1 사이의 값을 만들어 스케일을 표준화시키고, 이 유사도 값에서 역함수를 취해 '② 도출된 각도(theta)를 거릿값(0에서 90 사이의 값)으로 설정'한다. 그런 후 '③ 메타데이터 스페이스 상에 위치한 각 거릿값에 대해 유클리드 거리를 활용하여 최종적인 두 논문 간 거리를 제시'하였다. 그 이후 기준노드와 후보노드 두 논문 간의 거릿값을 도출하고 가까울수록 추천하는 방식을 채택하였다.

계량정보학에서의 저자-서지결합(Author Bibliographic Coupling; ABC), 동시인용분석, 단어 동시출현 분석의 방법을 적용하여, 본 연구에서는 저자-제목 결합, 키워드의 동시출현, 공저자 분석이 이뤄진다. 여기서 저자-서지결합(ABC)이란 서지결합(Bibliographic Coupling)에서 확장된 형태로(Gazni & Didegah, 2016), 두 저자가 인용한 논문의 저자가 일치하면, 두 저자의 연구 분야가 유사하다는 가정하에 이루어지는 분석이다(이재윤, 2008). 이러한 개념을 적용하여, 본 연구에서는 저자-서지결합분석에서의 '논문을 작성한 저자' 부분에는 '저자'를, 인용된 저자 부분에는 '제목'을 넣어 저자-제목 결합 유사도를 도출해보았다. 즉 저자-서지결합 분석에서는 인용한 저자들이 비슷할수록 인용을 수행한 저자끼리의 유사도가 높아진다는 의미를 가지는데, 본 연구에서의 저자-제목 결합에서는 제목의 요소가 비슷할수록 제목을 쓴 저자끼리의 유사도가 높아진다는 의미로 변형된다.

저자 혹은 키워드 단위의 유사도를 도출하는 방법은 기존 연구들의 동시인용 및 동시출현 분석의 방식을 활용하였다. 즉 두 저자 혹은 두 개의 키워드 간에 하나의 논문 내 동시 출현한 빈도를 계산하여 이를 통해 코사인 유사도를 도출하였다. 속성조합2의 '저자-공저' 유사도의 경우, 저자들이 같은 논문 내에서 서로 얼마나 공저자로 나타났는지에 따라 저자 간의 유사도가 도출되며, 속성조합4의 '키워드-동시출현' 유사도의 경우, 같은 논문 내에 키워드들이 얼마나 같이 출현했느냐에 따라 키워드 간의 유사도가 도출되는 형태이다.

각 속성조합 간의 코사인 유사도를 구한 후, 메타데이터 공간상 각 축 내 거리의 의미로 변환할 시 두 속성조합 간 세타(Θ)값을 도출하였다. 코사인 유사도 값은 두 노드 간에 가까울수록 1에 가까운 값을 가지는데, 거리의 개념은 0에 가까울수록 가깝다는 것이 직관적으로 와닿는다는 점에서 거리와 코사인 유사도는 반대의 값을 가진다고 볼 수 있다. 그러므로 코사인 유사도 상 세타값을 활용하는 방식을 취하여, 두 노드 간 가까울수록 0에 가깝게 변환하는 방식을 고안하였다.

각 거릿값(각도)의 생성 후 축 상에서 배치

가 되고, 이 거릿값을 가지는 축들이 직각으로 만나면 유클리드 공간이 생성된다. 이 공간 내에서 거리 공식을 활용하여 두 논문 사이의 거리를 계산하면, 최종적으로 두 논문 간의 거릿값, 즉 추천을 위한 값이 도출된다.

본 연구에서 메타데이터 공간 내 가장 많은 축을 가지는 추천조합은 RC03-01인데, 이를 예시로 거리 도출을 설명하면 다음과 같다. 이 추천조합은 '저자-제목(5-1)', '저자-공저(5-1)', '저자-논문간', '키워드-동시출현(3-3)', '키워드-논문간', '제목-논문간'의 속성조합으로 이루어져 있어, '저자'라는 속성에 총 11개의 축, '키워드'라는 속성에 총 10개 축, '제목'이라는 속성에 1개의 축을 가진다. 이렇게 형성된 최대 22개 축들을 모두 직각으로 두면, 하나의 논문은 22차원의 유클리드 공간 속에 하나의 좌표를 형성하게 된다. 저자 속성은 a~k축(11개 축)을, 키워드 속성은 l~u축(10개 축)을, 제목 속성은 v축(1개 축)을 가지게 되어, 논문1의 좌표는 (a_1, b_1, c_1, ... ,v_1)의 형태로 표현할 수 있으며 다른 논문 i의 경우 또한 (a_i, b_i, c_i, ... ,v_i) 형태의 좌표를 가질 것이다. 이러한 다른 좌표들과 기준이 되는 기준노드과의 거리를 비교하여 가장 가까이 위치한 논문(node)을 찾아내고 이를 추천하는 방식이다. 여기서 논문1(P1)과 논문2(P2)의 거리 측정 방식은 다음과 같다.

$$Distance(P1, P2) = \sqrt{(a_1-a_2)^2+(b_1-b_2)^2+(c_1-c_2)^2+...+(v_1-v_2)^2}$$

한편, 본 연구에서는 기준이 되는 노드의 좌표를 원점으로 보고, 다른 논문과의 거리는 코사인 유사도를 활용하여 도출해낸 세타값으로 나타내고자 한다. 그러므로 원점을 가지는 기준노드 P1과 속성조합별 P1과의 상대적인 거릿값(Θ)들을 가지는 후보노드 P2와의 최종적인 22차원에서의 거릿값 도출을 위해서는 다음과 같은 수식으로 변형이 가능하다.

$$Distance(P1, P2) = \sqrt{(0-\theta_a)^2+(0-\theta_b)^2+(0-\theta_c)^2+...+(0-\theta_v)^2}$$

$$= \sqrt{\theta_a{}^2+\theta_b{}^2+\theta_c{}^2+...+\theta_v{}^2}$$

where P1's coordinate is origin point

기준노드와 다른 노드 간의 거릿값만 알고 있으면 위의 수식과 같이 기준노드와 다른 노드와의 거리를 쉽게 도출해낼 수 있는 것이다. 이는 상대적인 거리 개념을 적용하여 계산된다. 여기에서의 상대적인 거리 개념은, 위의 예시와 같이 기준노드가 영점이 되도록 하여 두 노드 간 거리를 잰다는 의미가 된다.

각 후보노드들과 기준노드와의 상대적인 거리에 따라 추천 정도가 달라지는데, 거리의 개념이기에 값이 작을수록 근접한 학술 문헌으로 나타나며, 이를 활용해 가장 값이 낮은 상위 10개 혹은 5개 논문을 추천노드로 제시하는 방식인 것이다.

3.5 기준노드 설정 및 이용자 평가

3.5.1 기준노드 선정 및 타 시스템 추천 결과 수집

본 연구의 방법을 활용한 실험을 하고자 선정된 12개의 논문(기준노드)은 층화표집의 과

정을 거쳐 선정되었다. 저자의 특성이나 키워드(주제)의 특성, 제목 길이에 따른 특성을 살펴보면서 결과를 제시하고자, 특성별로 무작위 표집을 하였다. 무작위로 뽑힌 노드 중 본 연구에서 보고자 하는 특징을 잘 나타낸다고 판단되는 논문을 다시 꼼꼼히 확인 후 선정하는 절차를 거쳤다. 각 12개의 논문은 P1에서부터 P12까지 이름이 부여됐으며, 저자의 출현빈도에 따른 '연구 수행 다양성', 키워드 출현빈도에 따른 '주제 인기성', 제목 요소의 수에 따른 '제목 요소 정보량'의 특성을 매겼다. 이에 따라 기준노드를 분류하고 개별적 특성을 파악한 후, 비슷하거나 다른 성향을 가진 기준노드들끼리 묶어 이를 다음과 같이 분석하였다.

- P02 & P03: 대체로 주제 인기성을 가지거나 제목이 가지는 정보량이 적다는 특성이 있다. 두 논문 모두 '키워드' 메타데이터가 없었기 때문에, 제목 요소와 동일한 단어를 키워드에 넣었다.
- P04 & P08: 표본 내 키워드 출현빈도가 적은 특성이 있어, 비교적 적게 등장하는 연구주제를 지닌 논문들이다. 키워드 및 제목의 표본 내 출현빈도가 낮은 편이다. 한편 P04의 제목 요소 정보량은 많은 데 비해, P08은 적은 편이다.
- P07 & P08: '표본 내 다수의 공동저자를 가진 인기 있는 주제 vs. 세트 내 공동저자가 없는 관심이 적은 주제'로, 전반적으로 반대 특성을 가진다.
- P08 & P09: 두 논문 모두 표본 내에서 적게 다루는 주제를 가지고 있으며, 단일저자로 이루어져 있다는 특징을 가진다. 다만 차이점은 P09의 경우 표본 내에 동일 이름의 동일 소속 저자가 존재하며, P08은 표본 내에 단독으로 저자가 존재하여 이러한 차이로 인해 추천조합의 제안이 달라질 수 있을 것으로 보인다.
- P05 & P07: 저자의 연구 수행 다양성, 주제 인기성, 제목 요소의 정보량을 모두 충족하는 쌍이다.
- P02 & P12: 저자의 연구 수행 다양성과 주제 인기성을 확보했으나, 제목 요소의 정보량은 적은 편이다.
- P04 & P06: 저자는 표본 내 본 기준노드에서 한 번만 등장하고, 주제는 표본 내 적게 다루는 편이며, 제목 요소의 정보량은 많다.
- P01 & P03: 저자는 표본 내 본 기준노드에서 한 번만 등장하고, 제목의 정보량은 적지만 주제는 표본 내에서 인기가 있다.

3.5.2 타 시스템 추천 결과 수집

확보한 기준노드를 중심으로 타 시스템의 추천 결과를 수집하였다. 연구에서 제안하는 추천 결과에 대해 제대로 평가를 하려면, 타 추천시스템에 대한 비교도 필수적이다. 한국 논문에 대한 추천 평가이므로, 한국 논문의 검색을 제공하는 Google Scholar와 NDSL(2020년 실험을 수행할 당시 검색하였으며, 현재는 ScienceOn으로 통합되어 운영됨), DBpia를 중심으로 12개의 기준노드에 대한 추천 결과를 수집하였다.

학술데이터를 제공하는 웹사이트들의 추천 상황을 살펴보았을 때, 추천서비스를 제공하고 있는 곳은 그리 많지 않았다. NDSL의 경우 '이 논문과 함께 출판된 논문'이나 '저자의 다른 논문', '같은 권호 다른 논문', '인용한 논문(참고

문헌)' 등이 제공되고 있었고, 내용기반과 관련되어서는 NDSL과 DBpia에서 '같이 다운 받은 논문'이라든지 '이 논문과 함께 이용한 콘텐츠'와 같이 제시되어 있었다. 이에 NDSL과 DBpia는 다양한 방법론을 결합하여 논문을 추천하고 있음을 알 수 있었다. KISS나 KCI의 경우, '같은 권호 다른 논문'이나 '인용한 논문'에 대해서는 제공하고 있었으나, 추천논문은 나와 있지 않았다.

현황 분석 결과, 한국 논문에 대한 추천을 시도하고 있는 세 군데를 찾을 수 있었고, Google Scholar와 DBpia, NDSL이 그것이었다. Google Scholar의 경우 '관련 학술자료'로 나와 있었고, 가장 앞장에 제시되는 9~10개의 논문을 중심으로 수집하여 분석하였다. DBpia의 경우는 '추천 논문'으로 제시되었으며, 상위 5개가 추천되어 있어 이를 수집했다. NDSL은 '연관 논문'이라는 이름으로 서비스되고 있었으며, 상위 5개의 논문이 추천되어 있었기에 이를 수집하여 평가 및 비교 분석에 활용하였다.

기준노드 12개 중 수집한 추천을 제공했던 논문은 Google Scholar의 경우 11개, DBpia의 경우 8개, NDSL의 경우 4개였으며, 이를 확보하여 비교 분석에 활용하였다.

3.5.3 이용자 대상 실험 평가 및 분석

본 논문에서는 실험 결과에 대해 실이용자 중심의 주관적 평가를 시행하였다. 실제로 이용자가 추천 결과를 보며 얼마나 만족하였는지, 추천된 논문을 얼마나 이용하고 싶은지에 대한 점수를 확보했으며, 추천조합에 따라 추천된 순위를 보며 여러 추천조합에 대한 순위를 매기게 하여 이를 컨조인트 분석에 활용했다. 평가 결과는 전반적인 추천조합 및 타 시스템 대비 성능평가에도 활용하였다.

질적 평가를 위해서 세 명의 해당 분야 전공자인 이용자를 활용하였으며, 학술논문에 대한 추천을 평가하는 것이기에 학술논문에 대한 이해도가 높아야 하므로 평가자는 모두 평소 학술 문헌에 대한 정보를 추구해온 검색 시스템에 익숙한 박사급으로 구성하였다. 본 연구의 데이터가 전체적으로 사회복지학 분야와 경제학 분야에서 해당 키워드를 많이 다루고 있었고, 추천 결과로 관련 논문이 다수 등장을 하였기에 사회복지학 및 경제학 전공자로 구성하고자 하였다. 다음은 평가를 수행한 평가자 정보이다. 이용자1은 사회복지학 학・석・박사, 前 사회복지학과 교수이며, 이용자2는 경제학 학・석사, 사회복지학 학・박사과정, 경제사회 분야 연구원, 이용자3은 사회복지학 학사 및 문헌정보학 학・석사, 現 박사과정으로 선정하였다.

평가에 앞서 추천조합에 따라 다양한 추천 논문을 도출해냈고, 각 추천조합당 10위까지 도출이 되었다. 한 개의 기준노드에 50개의 추천조합이 있고, 평가자는 10위까지의 12개 논문에 대해 약 6,000개가량 논문 리스트를 검토해야 하는 상황이 되었다. 그러나 평가자들의 혼란을 방지하기 위해 해당하는 리스트 중 중복되는 것은 제외하고, 추천 순위 정보는 제거한 후, 블라인드 테스트 형식으로 추려진 리스트만을 가지고 선호도에 대한 평가를 하도록 하였다.

평가는 추천된 논문을 보면서, 기준노드와의 관련성을 확인하여 3단계로 활용 의지를 표명하였다. 이는 실제 영화 추천시스템에서 이용자들이 영화 평점을 매긴 결과를 활용하여 추

천을 수행하는 방식과 같이, 추천된 논문에 대한 이용자의 주관적인 선호도(평가/순서 매김 및 클릭 수)를 알아보고자 실시되는 것이다. 기준노드를 보고 있을 시 본 기법 및 타 시스템의 추천된 논문을 봤을 때 이를 '클릭'할 의향 정도에 따라 0~3점으로 표시하게 하였다. 평가 표현 시 0점은 '전혀 상관이 없는 논문이다', 1점은 '관련성이 낮은 편이어서 클릭하지 않을 것이다', 2점은 '관련성이 있으며 클릭해서 이용하고자 한다', 3점은 '관련성이 매우 높아서 이용하기 위해 클릭한다'로 이루어져, 0과 1점은 부적합으로, 2와 3점은 적합으로 나눈 후 이를 점수화하여 각 기법(시스템) 및 조합 간 평균을 비교 분석하였다.

3.5.4 평가 결과를 활용한 분석

선호도 분석을 위해서 이용자 평가 결과를 수집한 후, 평가자별로 부여한 평가값에 대해 평균값을 도출해냈다. 각 평가자는 시스템 이용자로서 참여했다는 데에 의의가 있으므로, 전문 학문 분야는 다르지만 각자 의견을 같은 무게로 인정하여 평가 결과는 산술평균으로 제시하였다. 그런 후 추천조합별로 각 평가된 논문들을 재배치하여, 이용자별 및 전체적인 추천조합 점수를 내서, 최종적으로 각 추천조합에 대한 선호도를 밝혀내고자 하였다.

추천된 논문 각각에 대한 평가자들의 선호도를 정리한 결과, 각 조합에 대한 평가자별 순위를 도출할 수 있었고, 이를 기반으로 표본 수 36개의 컨조인트 분석을 수행했다. 평가자들이 순위를 높게 매긴 조합이 무엇인지에 따라, 선호하는 속성조합의 구성을 도출해내어 각 논문 특성별 혹은 전반적으로 적합한 추천조합을 제안하고자 하였다.

각 추천조합 및 추천시스템 전반에 대한 선호도를 활용하여, 타 시스템 간의, 그리고 추천조합 간의 성능평가를 수행하였다. 본 연구에서는 근래 추천시스템 연구에서 많이 쓰이는 MRR(Mean Reciprocal Rank)과 MAP(Mean Average Precision)의 두 가지 성능평가 방법을 활용하였다.

4. 결과 논의

본 장에서는 제안한 추천기법으로 추천 리스트를 도출한 후 이용자들이 이를 평가하고, 그로부터 도출된 평가 점수를 활용하여 제안한 기법에 대해 분석하였다. 결과 분석에는 논문 및 추천조합 특성 분석, 각 논문과 조합별 상관계수 도출 및 컨조인트 분석, 서로 적합한 쌍의 도출, 타 시스템과의 비교 분석 결과를 제시하였는데, 이를 바탕으로 본 연구의 세 가지 연구 질문을 중심으로 한 논의를 수행하였다.

1) RQ1. 논문의 특성별로 효과적인 속성 및 추천조합이 달라지는가?

다음은 결과 분석 시 발견한 논문 쌍 및 적합한 조합에 대해 정리한 표이다.

〈표 3〉을 보면, 논문의 특성별로 어울리는 추천조합 및 추천조합군이 있음을 발견하였다. 이를 분석해보면, 대체로 '관련 세트 내 다른 저작이 있는 경우(연구 수행 다양성)'에는 오히려 이 저자 정보가 방해되므로, RC09군이 추천된다. 비슷하게는 '제목 정보량이 많은 경우' 또

〈표 3〉 논문 쌍과 적합한 추천조합군

논문 쌍(pair)		공통 특성	추천조합
P02	P03	키워드 메타데이터 없는 경우 제목 요소를 중복 사용 제목 요소의 수가 적음 관련 세트에서 주로 다루는 주제(주제 인기성 만족)	RC09군
P02	P07	상관계수 0.93 관련 세트 내 다른 저작이 있음(연구 수행 다양성 만족) 관련 세트에서 주로 다루는 주제(주제 인기성 만족)	
P02	P08	제목 요소의 정보량이 적은 경우	
P03	P11	제목 요소의 정보량이 적은 경우	
P04	P08	표본 내 키워드의 출현빈도가 적은 경우 관련 세트 내 다른 저작 없음 본 기법의 성능은 좋지 않은 편임	
P07	P09	상관계수 0.90 관련 세트 내 다른 저작이 있음(연구 수행 다양성 만족) - 저자의 특성이 오히려 추천에 방해	
P08	P09	단일저자, 표본 내 키워드의 출현빈도가 적은 경우 차이점: 연구의 다양성, 제목의 정보량	
P02	P09	상관계수 0.92 관련 세트 내에 다른 저작 있음(연구 수행 다양성 만족) 오히려 저자 정보를 활용하지 않는 것이 좋음	RC09군 RC07군
P07	P08	반대 특성을 가지는 쌍 저자 정보를 사용하지 않는 것이 좋음	
P03	P04	다수의 저자이며 관련 세트 내 다른 저작 없음	RC09군 RC03군
P05	P06	상관계수 0.82 제목의 정보량이 많은 경우 - 오히려 제목이 성능에 방해를 줄 수 있음	RC08군
P05	P10	상관계수 0.82 관련 세트 내 다른 저작 없음, 주제 인기성 높음 제목 정보량 많음	RC07군
P06	P10	상관계수 0.84 관련 세트 내 다른 저작 없음 제목 정보량 많음 키워드를 활용해야 성능을 높일 수 있음	RC07군 RC01군
P04	P11	상관계수 0.83 키워드 수 적음: 키워드 사용 않은 추천	RC05군
P04	P12	키워드 수 적음: 키워드 사용 않은 추천 본 기법의 성능은 좋지 않은 편임	
P11	P12	키워드 수 적음: 키워드 사용 않은 추천	
P03	P12	상관계수 0.88 관련 세트에서 주로 다루는 주제(주제 인기성 만족) 제목 정보량 적음	
P08	P10	저자 수 적음, 관련 세트 내 다른 저작 없음 논문 간 유사도 활용	RC02군
P01	P10	상관계수 0.87 단일저자, 관련 세트 내 다른 저작 없음, 주제 인기성 높음, 저자, 키워드, 제목 요소 적은 경우	RC01군

한 제목 속성이 오히려 방해를 줄 수 있어, 제목 속성이 빠진 추천조합이 좋은 성능을 보였다. 이처럼 특정 속성에 대한 정보를 다양하게 보유하고 있는 경우 더욱 좋은 결과를 보일 것이라 예상했었지만, 결과를 보면 다양하게 제공된 속성이 오히려 현재의 기준노드와는 전혀 다른 정보를 지닌 논문을 추천할 가능성도 가지고 있어 성능을 방해할 수 있음을 발견했다.

2) RQ2. 다양한 특성을 가진 논문들을 두루 고려했을 때, 전반적으로 추천 성능이 뛰어난 추천조합은 무엇인가?

전반적으로는 RC07군의 성능이 뛰어난 것으로 나타났다. RC07군의 경우 기준노드 사이에서 1위를 한 횟수가 비교적 적은 편이었기에, 상대적으로 좋은 성능의 추천조합군으로 RC09군이나 RC01군이 부각되었다. 하지만 컨조인트 분석의 최종 적합 조합으로 RC07군의 형태가 나타난 것으로 미루어보아, RC07군은 이용자들이 논문 특성에 상관없이 선호하는 RC군임을 알 수 있다. 그렇기에 향후 이 추천기법의 기본값을 설정하고자 할 때는 RC07군의 특징을 적용하면 대체로 높은 성능을 유지할 것으로 나타났다.

RC07군의 경우 '저자' 속성을 사용하지 않은 추천조합군이다. 실상 본 연구에서는 추천시스템상에서 성능을 높이는 데 저자의 역할이 매우 클 것이라 보았으나, 분석 결과 오히려 저자의 정보가 들어가지 않을 때 전반적으로 좋은 추천 결과를 낼 수도 있다는 것을 알게 되었다. 이는 '저자' 속성 자체가 지금 보고 있는 논문과 '주제 상 유사한' 논문을 찾아내는 것에 다소 방해가 된 것으로 파악된다. 실제로 데이터 및 결과를 보면, 저자의 정보가 많을수록 관련되지 않은 논문이 찾아지는 현상을 발견하였다. 동명이인의 문제가 발생하기도 하였고, 같은 사람이라도 공동연구 등과 같이 다루는 주제가 다양하면 이렇듯 오히려 저자 정보가 관련성 높은 문헌을 찾아내는 데 방해가 됨을 발견하였다. 이에 향후 성능을 높이고자 한다면, 저자 속성에 대한 깊은 이해와 저자 속성조합의 적합한 수 설정에 대한 고민이 필요하리라 본다.

3) RQ3. 기존의 추천시스템과 비교하였을 때 정보조직 및 계량정보학적 지식을 활용한 문헌정보학 기반 학술 문헌 추천기법의 효과적인 점과 한계는 무엇인가?

(1) 본 추천기법의 효과성을 높이는 요인

본 추천기법에 대한 평균 선호도 점수는 기준노드 P10이 가장 높아 본 기법으로 추천을 수행할 시 P10의 특징을 가진 논문들에 대해 좋은 성능을 보일 것으로 예상된다. P10을 살펴보면, 키워드 및 제목, 저자 구성에 중앙값의 특성이 있었고 표본 내 인기 있는 주제를 가지고 있었다. 이런 경우 본 논문의 기법이 적합하게 적용되어 추천 성능이 높게 나타남을 알 수 있다. 즉 여러 논문 가운데 상대적으로 키워드나 저자, 제목 등에 평범한 특성을 가진 논문의 경우 가장 좋은 결과를 냈다고 해석할 수 있다.

〈표 4〉에서 나타나듯, 타 추천시스템과 비교하여 본 기법이 이용자 선호도(pertinence)에서 1위를 차지(음영 표시)한 기준노드는 P02, P03, P05, P06, P07, P09, P10, P11이며, 대부분에서 1위를 하여 탁월한 결과를 보였다. 이

〈표 4〉 추천조합별 이용자 선호도(pertinence) 점수(평균)

	P01	P02	P03	P04	P05	P06	P07	P08	P09	P10	P11	P12	평균
본 기법	1.87	2.27	2.13	1.07	2.8	2.53	2.93	1.53	1.07	2.67	2.47	1.20	2.05
G.S.*	1.93	0.93	1	2.13	0.93		2.13	1.67	0.33	1.6	2.13	0.93	1.43
DBpia	1.33	0	1.73			0.87		0	1		1.27	1.20	0.93
NDSL	0									0.6	0.8	0.2	0.40
평균	**1.28**	**1.07**	**1.62**	**1.60**	**1.87**	**1.70**	**2.53**	**1.07**	**0.80**	**1.62**	**1.67**	**0.88**	**1.48**
본 기법 내 RC	01-1 01-5	09	09	05-2 외 2	08-2	08-3 08-4	09	02-1 02-2	09	01-3 외 3	05-4	05-4	

G.S.*: Google Scholar
값은 모두 이용자 평가 평균이며 3점 만점, 소수점 이하 반올림
'빈칸'은 해당 논문에 대해 추천된 논문이 없는 경우

중 P09의 경우는 네 시스템 전반에 걸쳐 추천 성능이 좋지 않았지만, 본 기법이 1위를 한 경우였다. P09에서 가장 성능이 좋았던 추천조합은 RC09인데, 제목 관련된 속성조합만 활용하는 군이며 이를 활용하면 적합도 및 이용자 선호도를 높일 수 있을 것이라 예상해볼 수 있다.

조합별 성능을 알아봤을 때, MRR과 MAP로 평가하니 P01과 P05, P07, P10, P11의 경우 본 추천기법을 통해 추천하였을 때 성능이 좋게 나왔다. 반면, P04, P08, P09, P12의 경우는 성능이 좋지 않은 편이었다. 성능이 높았던 노드의 경우, P11을 제외하고 키워드에서 드러나는 주제가 표본 내에서 많은 관심을 받으며 수행되는 경우였다. 한편 성능이 좋지 않은 경우는 P12를 제외하고 주제가 표본 내에서 많이 등장하지 않는 경우였다. 즉 표본 내에 얼마나 관련성 높은 연구가 많은지가 본 기법의 성능을 좌우하는 것으로 보여, 성능을 높이기 위해서는 광범위한 표본의 확보가 중요하다.

(2) 본 추천기법의 성능이 저하되는 요인

전반적인 성능이 좋지 않은 경우는 단일저자이면서 표본 내 다른 저작이 있어 저자의 정보가 방해되는 경우였다. 이를 해결하기 위해서는 기준노드의 저자 정보에 대해 전체적으로 파악한 후, 저자 정보 활용 여부를 결정하면 성능을 높일 수 있다.

키워드의 이슈 또한 있었다. 키워드에 대해 의미를 중시한 채로 접근하지 않고, 특정 키워드가 등장할 때 함께 등장한 키워드의 경우 관련 있는 키워드로 인식하여 제시하는 점이 오히려 결과에 방해로 작용하기도 하였다. 메타데이터를 작성함에 앞서 시스템 차원에서 full-text mining 및 TF-IDF 방식을 통한 통일화된 키워드 정제와 키워드 추출을 이룬다면, 이러한 문제는 대체로 해결될 수 있을 것이라 본다.

또한 일반적으로 잘 사용하지 않는 제목/키워드를 사용하는 논문이거나 대체로 모호하고 다중의 의미를 담고 있는 경우, 세부적인 주제에 맞는 논문을 찾아내기가 어려웠다. 모호하고 넓은 의미의 용어를 사용하는 경우는, 전문을 기반으로 하여 중심된 용어를 따로 추출하는 방식으로 접근하는 것이 나을 것이다. 메타데이터 기입 시 학술 문헌 수집처 자체에서 일괄적으로 통일시켜 키워드를 추출하는 기법을 활용한다면 성능이 나아질 수 있을 것이다.

(3) 본 연구의 차별성

본 연구는 여느 다른 추천시스템과는 다른 특성을 보이는데, 이는 '이용자의 프로필'을 보는 대신 '논문 자체(아이템)의 특성'에 집중했다는 것이다. 최근 추천시스템 연구의 추세는 인용 정보를 활용하거나, '이용자의 선호도', '이용자의 프로필' 등을 중점적으로 적용하는 방식을 가진다. 본 연구는 데이터 처리 시간이 상대적으로 적게 걸리는 메타데이터를 활용하고, 논문 자체의 다양한 특성을 살펴보며 적합한 추천 조합을 밝혀내는 데에 실험의 목적을 두었다. 이는 인용 정보 및 전문을 활용하여 추천을 수행할 시 시간과 저장공간이 많이 필요한 내용기반 추천과 비교된다. 이용자 프로필을 활용하는 것 또한 처리 시간이 상당히 필요한데, 개인마다 다르게 추천하는 알고리듬을 구축하는 것이기에 추천 결과의 다양화가 이뤄진다. 즉 사람이 n명이면 n개의 다양한 추천이 생기게 되는 것이다. 그리고 실이용자들의 경우, 로그인하지 않고 시스템을 사용하고자 하는 경우도 많아 이러한 프로필 기반의 추천을 구축해놓았으나 막상 활용이 어려운 경우도 다수 생긴다. 이에 본 연구는 이용자 개개인의 프로필에 맞춘 방식에 집중하는 것이 아닌 논문의 특정 특성들에 집중한 논문 추천을 시도하여, 이용자가 로그인하지 않아도 최적의 추천을 받을 방법을 고안해보았다. 즉 본 연구는 더 나은 성능의 문헌추천 시스템 생성의 기반을 마련하는 연구가 될 수 있다. 본 기법을 이용자의 로그인 전에 추천이 이루어지는 기본 방법으로 설정해놓고, 이용자가 로그인한 후에는 프로필 접근 방식을 추가하여 제시한다면 이용자들에게 보다 만족스러운 추천을 수행할 수 있으리라 본다.

(4) 본 연구의 한계 및 발전 방향

본 연구는 사전처리 시간이 많이 든다는 한계가 있다. 그러므로 속성조합상 축을 더 많이 사용하게 되는 조합과 덜 사용하는 조합의 성능이 비슷하다면, 시스템상 계산을 덜 사용하는 속성조합을 택하는 것이 안전하다. 이에 저자-제목 및 저자-공저 속성조합 중 (1-1)을 활용하는 것이 가장 좋은 선택임을 밝혀냈다. 이와 비슷하게는 저자(2-2)와 키워드-동시출현(3-3)을 활용한다면 제목을 사용하지 않아도 비슷한 성능을 보임을 밝혀냈고, 공저자 여부와 상관없이 비슷한 성능을 보이기도 했다(이는 컨조인트 분석에서의 결과와도 동일). 또한 키워드-동시출현을 활용할 시 키워드-논문 간 속성을 활용하지 않아도 활용한 것과 성능상 큰 차이가 없었다. 이에 메타데이터 공간상 축을 감소하여 속도와 성능을 높이고 시스템의 로드도 줄이는 방식을 고려하여 이를 적용해야 할 것이다.

본 연구 기법뿐 아니라 타 시스템에서도 추천 수행 성공률이 낮게 나온 논문들에 대해서 그 이유를 찾고자 노력할 필요가 있다. 본 연구에서 제안한 저자, 키워드, 제목 메타데이터 외에도 다양한 메타데이터를 활용한다면 추천 성능의 문제를 다소간 해소할 수 있을 것으로 보인다. 이를 위해서는 우선 추천이 어려운 논문들의 특징에 대한 파악이 이루어져야 할 것이고, 실험할 표본 내에 다양한 종류와 범위의 학술 문헌을 보유할 필요가 있다. 특히 외국에서 활발히 진행되고 있지만 아직은 국내로 도입되지 않은 연구라든지, 국외에서 막 시작되고 있는 연구의 경우 한국 논문만을 대상으로 추천을 수행하면 추천이 원활히 이루어지지 않아

한계가 있을 수밖에 없다. 그렇기에 논문 특성에 대한 철저한 분석과 다양한 표본의 확보를 지속적으로 시도하여 전반적인 추천 성능을 높이기 위한 노력을 기울일 필요가 있다.

5. 결 론

본 연구는, 국내외적으로 대부분 컴퓨터공학 분야에서 연구되고 있는 추천시스템을 문헌정보학적인 시각으로 접근하여 방법론을 제안했다는 점에서 새로운 시도였다고 본다. 특히 코사인 유사도의 개념 및 '계량정보학적 기법'과 디지털 정보조직 시 주로 다루게 되는 '메타데이터'를 활용하여 문헌에 대한 추천서비스를 고안해보았는데, 이처럼 문헌정보학의 본질을 최대한 살린 추천기법을 제안했다는 점과 문헌정보학적인 접근법을 통해서도 충분히 추천 성능을 높일 가능성을 발견했다는 점에서 본 연구의 의의가 있다.

특히 본 연구에서 제안한 추천 결과의 성능이 우수했음에 주목해야 한다. 다른 학술 문헌 추천시스템(Google Scholar, DBpia, NDSL)과의 수행결과를 비교한 결과, 전반적으로 본 방법의 추천 성능이 1위를 기록하였다. 1위를 하지 못하거나 낮은 점수를 받은 경우도 충분치 못한 표본의 문제로 인해 세트 내에 적합한 논문이 부족하거나, 타 시스템에서도 추천 수행결과가 좋지 않은 논문일 때였던 것으로 미루어보아, 성능 자체의 문제라 보기 어려웠다. 그러므로 논문 정보 파악을 통해 서로 간 유사도를 다차원 상에 좌표화 하여 이들 간의 거리를 산출하고, 거릿값이라는 하나의 수치로 기준노드와 추천노드와의 거리를 계측해낸 이 방법은 추천 결과 성능에서 기존의 방법에 비해 매우 뛰어나게 작동함을 확인할 수 있었다.

향후 본 연구를 활용하여 실제 시스템으로 구현할 시, 메타데이터의 활용 속성에 대한 정리 및 정제에서 시간이 다소 걸릴 것으로 판단된다. 하지만 이러한 문제는 시스템 자체에 메타데이터를 입력하기 전 단계에서 미리 메타데이터 기입에 대한 표준화를 이루어, 오류 없이 데이터를 저장해 놓으면 해결될 수 있다. 또한 시스템상에 용어집 확보를 하는 것이 필수적이다. 특히 키워드의 경우 동일한 쓰임을 가지는 용어임에도 다르게 표현하고 있는 경우도 많아 다른 용어로 인식하게 되는 문제도 생겼다. 이는 결과에 오류가 생기게 되어 성능을 떨어뜨리는 요소로 작용한다. 이런 경우 외국어를 다룬 용어집이나 한글의 유의어 사전을 활용하여, 같은 뜻을 가진 단어가 다양한 형태로 표현되는 경우 이들을 동일어로 인식하게끔 해야 할 것이다. 상기에 언급한 부분들은 어떠한 추천시스템으로 추천을 수행하더라도 마찬가지로 중요한 부분이므로, 검색 효율 및 검색 성능을 높이기 위해서는 꼭 정리해야 할 필수적인 전처리 과정이라 볼 수 있다.

RC07군 및 RC09군에서 좋은 성능을 보였는데, 이 추천조합은 다른 조합들에 비해 상대적으로 간단하게 이루어져 있다는 특징이 있다. 이에 후속 연구에서는 간단한 방식의 추천을 기본 설정으로 두고서, 점차 복잡한 조합을 추가해나가면서 성능을 비교해볼 수 있을 것이다. 즉 시스템 내 각 속성을 자동 색인하여 통합색인을 만든 후 이를 기본 설정으로 두고 시작하는 방식이다. 이는 기존 색인 및 검색 모듈을

활용하는 것이므로, 본 연구의 사전처리 시간 문제를 해결할 방법이 될 수 있을 것이다.

본 연구에서는 제목 정보량이 많은 경우 제목 속성이 오히려 방해되는 경우, 관련 세트 내 다른 저작 여부로 인해 저자의 정보가 방해되는 경우, 다양한 속성이 오히려 전혀 다른 주제의 논문을 추천하는 경우 등을 발견하였다. 이는 거릿값 계산에 0에서 90 사이의 세타값을 활용하여 폭넓은 분포로 인해 편차가 극단적으로 나타나 일어난 일이라고도 볼 수 있다. 이에 향후 연구에서는 평균이 0이고 표준편차가 1인 피어슨 상관계수를 세타값 대신 산출하여, 각 값 간 편차를 줄인 유클리드 거리를 적용해볼 수 있을 것이다. 이처럼 거릿값 자체를 변환하여 이로 인해 달라지는 추천 결과를 비교 확인하는 것도 전반적인 성능을 높이는 데에 의의가 있다고 본다.

실험에 활용한 표본의 경우, 국내외 전체 학술 문헌을 사용했더라면 전반적인 성능을 살펴볼 수 있었을 것이며, 다른 시스템과의 비교에서도 더욱 유리하게 작용했을 것으로 보인다. 다른 시스템에서는 좋은 평점으로 추천되었던 논문이 실험 표본 상에는 없는 경우가 다수였기 때문이다. 실험에 사용했던 데이터가 '격차(불평등)' 관련한 데이터로 한정되어 있었기에, 격차나 불평등에 대해 다루고 있지 않으나 기준노드와 관련성이 높은 논문은 추천되지 못하였다. 본 연구에서 표본을 한정시킨 이유는 실험의 속도를 높이고 세분화된 분야에 대해 상세히 분석하고자 함이었으나, 결과적으로는 한계로 작용하게 되었다. 또한 본 연구는 KCI에서 제공하는 국내 연구 메타데이터만 다루고 있었기에, 관련이 있는 국외 연구물에 대해서는 추천해낼 수 없었다. 그래서 향후 연구를 수행한다면, 전 세계의 학술 문헌을 전체적으로 활용하여 이를 기반으로 추천을 수행하고, 다른 시스템과 동일 선상에서 비교 및 분석을 하여 본 기법의 발전 가능성을 구체적이고도 명료하게 살펴보아야 할 것이다.

본 연구에서는 전반적으로 저자, 제목, 키워드의 세 가지 메타데이터 속성을 활용해서 연구를 수행했지만, 향후 더욱 다양한 메타데이터 속성을 활용한다면 사회의 기타 여러 추천 분야에 도움을 줄 방법으로 발전할 수 있을 것이다. 더불어 이용자의 의견 및 선호도를 시스템상에 조합한다면, 진정한 의미의 혼합형(hybrid) 기법으로서 발전할 수도 있을 것이며 성능 또한 높일 수 있을 것이다. 이처럼 다차원 메타데이터 공간상에 적용할 수 있는 다양한 정보 형태들을 탐색하고, 각 자원의 메타데이터 속에 숨겨져 있는 양질의 정보들과 최상의 조합을 발견해낸다면, 인간의 지적 구조에 가깝게 작용하는 시스템, 더 나아가서는 AI의 연구영역에도 도움을 줄 수 있는 시스템으로 확장하여 활용될 수 있을 것으로 기대된다.

참 고 문 헌

고영만, 송민선, 이승준 (2015). 한국학술지인용색인(KCI)의 인문학, 사회과학, 예술체육 분야 저자키워드의 의미적 관계 유형 최적화 연구. 한국문헌정보학회지, 49(1), 45-67. http://dx.doi.org/10.4275/KSLIS.2015.49.1.045

박대우, 고인수, 이낙선, 한경석 (2020). 빅데이터 기반 도서추천시스템 구축을 위한 아키텍처에 관한 연구. 한국IT정책경영학회 논문지, 12(1), 1559-1565.

여운동, 박현우, 권영일, 박영욱 (2010). 연구논문 추천시스템의 전자도서관 적용방안. 한국콘텐츠학회 논문지, 10(11), 10-19. http://dx.doi.org/10.5392/JKCA.2010.10.11.010

원재상 (2020). 문맥을 고려한 논문 추천 시스템. 한국정보과학회 학술발표논문집, 1620-1622.

이재윤 (2008). 서지적 저자결합분석: 연구동향 분석을 위한 새로운 접근. 정보관리학회지, 25(1), 173-190. http://dx.doi.org/10.3743/KOSIM.2008.25.1.173

임윤정, 송규원, 조민상, 정현준 (2021). 학술 빅데이터 기반 분야별 지능형 전문가 추천 시스템. 한국정보과학회 학술발표논문집, 111-113.

한국연구재단 (2020). 한국학술지인용색인(KCI) DB 정보. 출처: https://www.kci.go.kr/

Adomavicius, G. & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE transactions on knowledge and data engineering, 17(6), 734-749. http://dx.doi.org/10.1109/TKDE.2005.99

Ahmad, S. & Afzal, M. T. (2017). Combining co-citation and metadata for recommending more related papers. 2017 International Conference on Frontiers of Information Technology (FIT). IEEE, 218-222. http://dx.doi.org/10.1109/FIT.2017.00046

Ahmad, S. & Afzal, M. T. (2020). Combining metadata and co-citations for recommending related papers. Turkish Journal of Electrical Engineering & Computer Sciences, 28(3), 1519-1534. https://doi.org/10.3906/elk-1908-19

Alshareef, A. M. (2019). Academic Recommendation System Based on the Similarity Learning of the Citation Network Using Citation Impact. Doctoral dissertation, University of Ottawa. http://dx.doi.org/10.20381/ruor-23359

Beel, J. & Gipp, B. (2009). Google Scholar's ranking algorithm: the impact of citation counts (an empirical study). In 2009 third international conference on research challenges in information science, 439-446. IEEE. http://dx.doi.org/110.1109/RCIS.2009.5089308

Beel, J., Gipp, B., Langer, S., & Breitinger, C. (2016). Paper recommender systems: a literature survey. International Journal on Digital Libraries, 17(4), 305-338.

https://doi.org/10.1007/s00799-015-0156-0

Chen, T. T. & Lee, M. (2018). Research paper recommender systems on big scholarly data. In Pacific Rim Knowledge Acquisition Workshop, 251-260. Springer, Cham. https://doi.org/10.1007/978-3-319-97289-3_20

Deldjoo, Y., Dacrema, M. F., Constantin, M. G., Eghbal-Zadeh, H., Cereda, S., Schedl, M., Ionescu, B., & Cremonesi, P. (2019). Movie genome: alleviating new item cold start in movie recommendation. User Modeling and User-Adapted Interaction, 29(2), 291-343. https://doi.org/10.1007/s11257-019-09221-y

Gazni, A. & Didegah, F. (2016). The relationship between authors' bibliographic coupling and citation exchange: analyzing disciplinary differences. Scientometrics, 107(2), 609-626. https://doi.org/10.1007/s11192-016-1856-y

Khan, S., Liu, X., Shakil, K. A., & Alam, M. (2017). A survey on scholarly data: From big data perspective. Information Processing & Management, 53(4), 923-944. https://doi.org/10.1016/j.ipm.2017.03.006

Liling, L. I. U. (2019). Summary of recommendation system development. In Journal of Physics: Conference Series 1187(5), 052044. IOP Publishing. http://dx.doi.org/10.1088/1742-6596/1187/5/052044

Morris, S. A. & Yen, G. G. (2004). Crossmaps: Visualization of overlapping relationships in collections of journal papers. Proceedings of the National Academy of Sciences, 101(Suppl. 1), 5291-5296. https://doi.org/10.1073/pnas.0307604100

Nazir, S., Asif, M., & Ahmad, S. (2020). Exploring the Proportion of Content Represented by the Metadata of Research Articles. In 2020 3rd International Conference on Advancements in Computational Sciences (ICACS), 1-7. IEEE. https://doi.org/10.1109/ICACS47775.2020.9055955

Waheed, W., Imran, M., Raza, B., Malik, A. K., & Khattak, H. A. (2019). A hybrid approach toward research paper recommendation using centrality measures and author ranking. IEEE Access 7, 33145-33158. https://doi.org/10.1109/ACCESS.2019.2900520

Williams, K., Wu, J., Choudhury, S. R., Khabsa, M., & Giles, C. L. (2014). Scholarly big data information extraction and integration in the citeseer χ digital library. In 2014 IEEE 30th International Conference on Data Engineering Workshops, 68-73. IEEE. https://doi.org/10.1109/ICDEW.2014.6818305

Xia, F., Wang, W., Bekele, T. M., & Liu, H. (2017). Big scholarly data: A survey. IEEE Transactions on Big Data, 3(1), 18-35. https://doi.org/10.1109/TBDATA.2016.2641460

Yang, S., Han, R., Wolfram, D., & Zhao, Y. (2016). Visualizing the intellectual structure of information science (2006-2015): Introducing author keyword coupling analysis. Journal of informetrics, 10(1), 132-150. https://doi.org/10.1016/j.joi.2015.12.003

• 국문 참고문헌에 대한 영문 표기
(English translation of references written in Korean)

Im, YunJeong, Song, Gyuwon, Cho, MinSang, & Jung, HyunJung (2021). Intelligent export recommendation system based on academic bigdata. Proceedings of the Korean Information Science Society Conference, 111-113.

Ko, Young Man, Song, Min-Sun, & Lee, Seung-Jun (2015). A study on the optimization of semantic relation of author keywords in humanities, social sciences, and art and sport of the Korea Citation Index (KCI). Journal of the Korean Society for Library and Information Science, 49(1), 45-67. http://dx.doi.org/10.4275/KSLIS.2015.49.1.045

Lee, Jae-Yun (2008). Bibliographic author coupling analysis: a new methodological approach for identifying research trends. Journal of the Korea Society for Information Management, 25(1), 173-190. http://dx.doi.org/10.3743/KOSIM.2008.25.1.173

National Research Foundation of Korea (2020). Korea Citation Index(KCI) DB Information. Available: https://www.kci.go.kr/

Park, Dae-Woo, Koh, In Soo, Lee, Nak-Son, & Han, Kyeong-Seok (2020). A study on architecture for bigdata-based book curation system. Jounal of The Korea Society of Information Technology Policy & Management, 12(1), 1559-1565.

Won, Jaesang. (2020). Context-aware recommendation system for literature. Proceedings of the Korean Information Science Society Conference, 1620-1622.

Yeo, Woon-Dong, Park, Hyun-Woo, Kwon, Young-Il, & Park, Young-Wook (2010). Application of research paper recommender system to digital library. The Journal of the Korea Contents Association, 10(11), 10-19. http://dx.doi.org/10.5392/JKCA.2010.10.11.010