DOI : 10.5392/JKCA.2010.10.11.010

연구논문 추천시스템의 전자도서관 적용방안

Application of Research Paper Recommender System to Digital Library

여운동, 박현우, 권영일, 박영욱
한국과학기술정보연구원

Woon-Dong Yeo(wdyeo@kisti.re.kr), Hyun-Woo Park(hpark@kisti.re.kr),
Young-Il Kwon(ylkwn@kisti.re.kr), Young-Wook Park(ywpark@kisti.re.kr)

요약

컴퓨터와 웹의 발달은 사람들이 이용할 수 있는 정보의 양을 급격히 늘렸으며, 이로 인해 추천시스템에 대한 수요가 증가하고 있다. 전자도서관에서도 다른 분야와 마찬가지로 개인화 및 추천시스템에 대한 연구가 중요한데, 연구논문 추천시스템에 대한 연구는 극히 제한적으로 이뤄지고 있고, 국내에서는 거의 찾아보기 어려울 정도이다. 본 논문에서는 국내외에서 수행된 추천시스템에 대한 연구를 조사분석하고, 이를 토대로 전자도서관 연구논문 추천시스템 구축방안을 KISTI NDSL을 중심으로 제안한다. 현재 NDSL에서 제공하는 알리미서비스를 암묵적 방식으로 바꾸어서 이용자의 프로파일을 구축할 것과 이용자 및 메모리 기반의 협업 필터링을 병행하여 내용기반의 필터링이 가지는 연구논문 추천에서 신규성이 부족한 단점을 보완할 것을 제안한다. 또한 두 기법을 함께 사용하는 방식과 온톨로지와 분할방식에 의한 필터링을 이용하여 추천 만족도를 높이는 방식에 대해서도 제안한다.

■ 중심어 : | 추천시스템 | 개인화서비스 | 협업 필터링 | 전자도서관 |

Abstract

The progress of computers and Web has given rise to a rapid increase of the quantity of the useful information, which is making the demand of recommender systems widely expanding. Like in other domains, a recommender system in a digital library is important, but there are only a few studies about the recommender system of research papers, Moreover none is there in korea to our knowledge. In the paper, we seek for a way to develop the NDSL recommender system of research papers based on the survey of related studies. We conclude that NDSL needs to modify the way to collect user's interests from explicit to implicit method, and to use user-based and memory-based collaborative filtering mixed with contents-based filtering(CF). We also suggest the method to mix two filterings and the use of personal ontology to improve user satisfaction.

■ keyword : | Recommender System | Personalisation | Collaborative Filtering | Digital Library |

I. 서 론

컴퓨터와 웹의 발달은 유용한 정보와 유용하지 못한 정보를 동시에 급격히 늘어나게 하였다. 이로 인해 키워드를 이용한 단순검색은 이용자를 더 이상 만족시키지 못하고 있으며, 미처 발견하지 못한 중요한 정보에

접수번호 : #101014-002
접수일자 : 2010년 10월 14일
심사완료일 : 2010년 11월 25일
교신저자 : 박영욱, e-mail : ywpark@kisti.re.kr

대한 불안감 및 갈증이 증가시키고 있다. 정보 이용자는 대부분 자기가 원하는 것이 무언이지 정확히 알지 못하거나 그것을 한두 개의 키워드로 표현할 수 없으며, 또한 기존에 경험하지 못했던 것에 대해 막연한 흥미를 가진다. 이럴 경우 누군가 자기가 찾는 것을 상황에 맞게 “추천”해 준다면 이용자들은 한결 편하게 정보를 선별하여 이용할 수 있을 것이다.

추천은 정보나 서비스가 개인이나 특정집단의 요구에 맞게 재단된 개인화(personalisation)의 가시적 형태를 말하며 이러한 기능을 하는 시스템을 추천시스템(recommender system)이라고 한다[5]. 추천시스템은 이용자가 눈여겨 봐야할 것을 선별해줄 뿐만 아니라 그럴 필요가 없는 것을 복잡한 정보네트워크를 탐색하여 걸러 주는 기능도 한다. 이미 많은 분야에서 추천시스템이 활용되고 있으며 성공한 시스템으로 평가받고 있다. 대표적인 분야로는 뉴스[16], 영화[18], 음악[30], 유머[19] 등이다. 최근에는 미국의 온라인 DVD 대여업체인 Netflix에서 2011년까지 자체개발한 추천시스템보다 10% 향상된 성능을 가진 알고리즘을 개발하면 100만 달러의 상금을 지급하겠다는 조건을 내걸기도 할 정도로 추천시스템의 수요가 증가하고 있다.

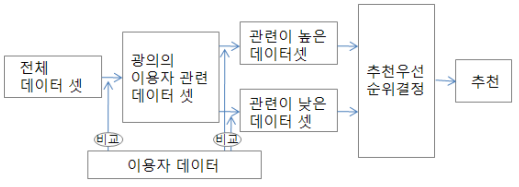

그림 1. 일반적인 추천 프로세스

전자도서관에서도 다른 분야와 마찬가지로 개인화 및 추천시스템에 대한 연구가 중요한 과제이며, 실제 이용자를 대상으로 서비스를 하는 도서관들이 있다. 그러나 Mylibrary[14], ACM Digital Library[18] 등 대부분의 전자도서관에 적용된 많은 개인화 응용기술은 기본적인 개인화 시스템이 적용된 알림 서비스로서 개인의 관심 주제에 부합하는 논문이나 책이 새롭게 들어왔다는 것을 개인에게 통보해 주는 수준이다[5][12]. 이렇게 전자도서관에 적용된 추천시스템 수준이 높지 않은 이유는 전자도서관과 관련된 연구와 개발의 대부분이 Dublin Core와 같은 데이터 표준화, 정보처리상의 상호운용성, 디지털 권리보호를 위한 기술, 메타데이터의 자동생성, 디지털 객체 식별 등 디지털 프로세스에 집중되어 왔기 때문이다. WoS나 SCOPUS에 등재된 논문에서도 다른 분야에 비해 전자도서관 추천시스템에 대한 논문의 수가 현격이 적게 검색된다. 특히 연구논문 추천시스템에 대한 연구는 극히 제한적으로 이뤄지고 있고, 국내에서는 거의 찾아보기 어려울 정도이다.

미국 과학재단(National Science Foundation)에 의하면 1999년에 전 세계적으로 1,900개 이상의 저널에서 50만개가 넘는 연구논문이 발행되었으며, 1986년 이후 그 수가 매년 1%씩 증가하고 있다고 한다[24]. 발표되는 논문의 수가 많아진다는 것은 참고할 논문이 많아진다는 점에서 연구자들에게 행복한 소식이 아닐 수 없다. 하지만 한편으로는 자신의 읽기 목적과 좀 더 잘 맞는 논문을 찾기 위해 더 많은 시간을 투자해야 한다는 것을 의미하기도 한다. 또한 연구자들은 논문의 우수성에 대한 검증도 함께 수행해야 하는 부담을 더 가질 수밖에 없을 것이다. 추천시스템은 논문의 읽기 목적 뿐 아니라 논문의 우수성도 함께 고려하여 추천을 할 수 있다.

논문정보 검색에서 추천시스템 도입이 필요한 또 다른 이유 중에 하나는 연구자들의 검색능력이 그다지 높지 않다는 것이다. 심지어 경험 많은 연구자조차도 데이터베이스에서 검색을 잘 하지 못하거나 검색을 통해 나온 논문을 최근기준으로 한 두 개의 논문만 보거나 검색결과 첫 페이지만을 찾는 수준에 그친다[22]. 연구자들은 질문식을 세밀하게 검토하여 좀 더 나은 검색식을 만들거나 관련성이 낮은 논문을 쉽게 지나치지도 못하고 있다. 정보공학자가 생각하는 재현율이나 정확성은 일반 연구자의 검색에서는 큰 의미가 없어 보인다. 논문정보를 제공하는 데이터베이스를 마치 일반 검색을 전문으로 하는 구글과 같은 식으로 사용하고 있는 것이다.

기업은 궁극적으로 이용자들의 기호에 잘 맞는 서비스를 눈에 띄도록 추천하여 상품판매수를 늘리기 위해서 추천시스템을 도입하고 있다. 이에 반해 NDSL(www.ndsl.kr)과 같은 국가 전자도서관의 논문

추천시스템은 이용자들이 많은 논문을 읽게 한다는 것보다는 연구목적과 맞는 논문을 추천하여 연구자들의 연구 능력과 효율 향상에 기여하는 데에 궁극적인 목적이 있다. 많은 논문을 읽게 하기 보다는 오히려 더 적은 논문을 읽고도 원하는 목적을 달성할 수 있게 해 주어야 한다. 이 목적은 미국 Arrowsmith 프로젝트[22][23]와 같이 연구자가 풀어야 하는 난제에 대한 해결책(논문)을 제시하는 것과 연구자가 앞으로 수행해야만 하는 유망한 기술을 추천하는 것 등을 통해 한층 더 높은 수준으로 달성할 수 있다. 연구논문의 추천이 상품이나 서비스 추천과 다른 차원의 중요성에서 다루어져야 하는 이유가 여기에 있다.

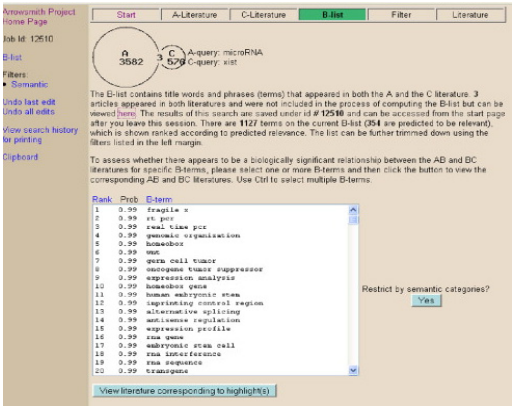

그림 2. Arrowsmith 검색화면[23]

그동안 국내 연구논문 추천시스템에 대한 연구는 세부 알고리즘을 제안하고 그것에 관해 기존 보다 성능이 우수하다는 것을 증명하는 것이 대부분이었다. 추천 결과에 대한 이용자 만족도는 이용자 직업/경력/연령 등의 분포와 이용 목적, 서비스 환경, 부가서비스, 데이터 특성 등이 복합적으로 작용하기 때문에 시스템 개발에 앞서 기존 연구결과를 폭넓게 검토하는 것이 필요하다. 그렇지만 필요에 의해서 개발을 진행하는 경우 오히려 기존 연구에 대한 조사분석 과정을 소홀히 하는 경우가 많다. 본 논문에서는 국내외에서 수행된 연구논문 추천시스템 연구의 결과를 비교분석하여 향후 국내전자도서관 연구논문 추천시스템 개발방향을 제안하고자 한다. 국내에서 연구논문을 서비스하는 대표적인 전자도서관은 국회도서관, 한국교육학술정보원 RISS, 한국과학기술정보연구원 NDSL 등이 있다. 본 논문에서는 특허정보 등 다양한 연구정보를 함께 제공하는 NDSL을 중심으로 연구논문 추천시스템 구축 방안을 모색한다.

본 논문의 구성은 다음과 같다. Ⅱ장에서는 추천시스템 기법과 각 기법이 갖는 장단점을 살펴본다. Ⅲ장에서는 연구논문 추천시스템에 적용 가능한 추천기법과 각 기법에서 발생할 수 있는 문제점을 해결할 수 있는 방안을 모색한다. 끝으로 Ⅳ장에서는 본 논문의 결론과 향후 연구방향을 기술한다.

Ⅱ. 추천기법

추천기법의 일반적인 분류는 [표 1]과 같다. 유사성의 대상이 내용기반인지 이용자 간의 협업 기반인지에 따라서 내용기반과 협업 필터링으로 나뉘고, 이용자가 이용하지 않은 상품이나 정보(이하 아이템으로 통칭함)에 대해서 선호도를 추정할 때 이용자의 과거 선호도를 기반으로 하는지 모델을 기반으로 하는지에 따라서 메모리기반과 모델기반 필터링으로 나눌 수 있다[7][11].

본 장에서는 각 추천알고리즘에 대해 간단히 살펴본다. 본 논문은 NDSL 추천시스템 구축 방안에 대한 논리적 접근을 연구의 목적으로 하기 때문에 본 논문에서는 추천시스템 알고리즘은 기본 원리만 살펴본다.

표 1. 추천시스템 분류[11]

추천방법	주로 사용되는 기법	
	메모리기반	모델기반
내용기반 필터링	·TF-IDF ·Clustering	·Bayesian classifiers ·Clustering ·Decision Tree ·Artificial neural networks
협업 필터링	·Nearest neighborhood ·Clustering ·Graph theory	·Bayesian classifiers ·Clustering ·Artificial neural networks ·Linear regression ·Probabilistic models
혼합형 필터링	·Linear combination of predicted ratings ·Various voting schemes ·Incorporating one component as a part of the heuristic for the other	·Incorporating one component as a part of the model for the other ·Building one unifying model

1. 내용기반 필터링과 협업 필터링

내용기반 필터링 기법은 정보 검색이나 정보 필터링 연구에서 자연적으로 발전하였다. 따라서 내용기반 필터링은 대부분 아이템을 추천하기 위해 아이템의 내용과 이용자의 정보요구간의 유사도를 측정하고 그 결과를 순위화하여 보여준다[3]. 이 기법은 정보검색에 기반을 두고 있으므로 이 분야에서 자주 사용되는 기법인 가중치 기법, 적합성 피드백, 불리안 검색, 확률검색 모형 등을 활용한다.

내용기반의 필터링은 영화, 음악과 같이 텍스트와 무관한 주제에 대해서는 적용이 불가능하고, 글의 스타일이나 수준, 저자의 권위 등 이용자들이 아이템에 대해 가질 수 있는 주관적인 판단에 따라 유사성을 군집할 수 없으며, 단지 과거에 관련이 있는 아이템만을 추천할 뿐이지 내용에 포함된 단어는 달라도 관련된 아이템인 경우에도 추천이 불가능하다는 단점이 있다[20].

협업 필터링은 추천시스템에서 자장 흔히 쓰이는 기법이다. 협업 필터링을 이용한 추천시스템은 이용자의 선호도를 수집하여 데이터베이스를 구축하고, 특정 이용자와 유사한 취향이나 정보요구를 갖는 이웃들을 데이터베이스에서 찾아내어 이들이 좋아하는 아이템을 이용자에게 추천한다[3].

협업 필터링은 해결해야할 몇 가지 중요한 도전과제가 있다[25][11]. 첫 번째는 결측치 및 희소성의 문제로, 일반적으로 이용자들이 전체 아이템(정보)에 비해 극히 일부에 대해서만 자기의 선호도를 표시하게 된다. 이것은 결국 이용자 사이의 유사성 계산에 좋지 못한 결과를 가져 온다. 따라서 이 경우에는 디폴트값과 같은 특수한 데이터 처리를 필요로 한다.

협업 필터링이 가지는 또 다른 문제는 이용자의 선호도와 이용자 사이의 유사성의 불일치이다. 이것은 이용자 간에 존재하는 선호도를 표시하는 습관(패턴)의 차이에서 비롯된다. Correlation 계수가 Cosine 계수보다 이용자 유사성 측정이 더 우수하게 나타나는 이유도 여기에 있다. 그러므로 이용자의 선호도를 이용자의 습관을 고려하여 상대적으로 계산할 수 있는 데이터 처리가 있어야 한다(연구논문 추천시스템에서는 인용데이터나 이용자의 클릭 혹은 다운로드를 기준으로 선호도를 측정하게 된다. 그러므로 이용자의 선호도 표시 패턴은 일정하게 1이나 0로 나타나므로 Correlation 계수보다는 Cosine 계수의 사용이 더 적합할 것으로 판단된다).

시스템에 이용자 선호도가 없거나 부족하여 이용자와 유사한 패턴의 선호도를 가지는 이웃을 찾을 수 없을 때 발생하는 새로운 시스템 문제(New system cold star problem), 새로운 이용자가 시스템에 가입 하였을 때 이용자에 대한 정보가 없어서 이웃을 찾을 수 없게 되는 새로운 이용자 문제(New user colds star problem), 신규 아이템이 시스템에 유입되었을 때 시스템에 대해 평가를 한 이용자가 없기 때문에 발생하는 새로운 아이템 문제(New item cold star problem)도 협업필터에서 해결해야 할 과제이다[6].

마지막으로 협업 필터링의 해결해야 할 중요한 과제는 확장성(Scalability)과 성능향상을 동시에 달성해야 하는 것이다.

협업 필터링 알고리즘은 실시간으로 대규모의 잠재적 이웃을 검색할 수 있어야 한다[9]. 만약 아이템 정보를 많이 가진 이용자들이 존재한다면 검색시간이 더욱 늘어날 것이다. 설상가상으로 이용자의 만족도를 높이기 위해서는 계산량이 좀 더 많은 협업 필터링 알고리즘을 사용해야 한다. 이는 확장성과 성능향상이라는 두 가지가 상호 모순으로 작용하게 한다.

2. 메모리기반 필터링과 모델기반 필터링

메모리기반 알고리즘은 이용자가 아이템에 대해 표

그림 3. NDSL에서 제공하는 관련문헌 서비스

시한 선호도를 바탕으로 현재 테스트 이용자와 유사한 패턴을 보이는 이웃을 훈련 데이터베이스에서 찾고, 테스트 이용자와 이웃과의 연관성 및 이웃이 표시한 선호도에 비례하여 테스트 이용자가 이용경험이 없는 아이템에 대해 선호도를 추정하여 높은 값을 보이는 아이템을 추천하는 방식이다.

모델기반 알고리즘은 훈련 데이터베이스에서 이용자의 선호도 패턴을 바탕으로 이용자를 작은 단위의 그룹으로 분류한 뒤, 테스트 이용자와 가장 가까운 그룹을 선호도 패턴으로 찾고, 그룹의 선호도에서 테스트 이용자의 선호도를 추산해 낸다.

모델기반 알고리즘은 복잡한 계산을 필요로 하기 때문에 메모리기반 알고리즘보다 오프라인에서 수행되는 이용자 간의 유사도 계산량이 많다는 단점이 있으나, 온라인에서는 모델의 프로파일만 저장하면 되므로 전체 이용자의 프로파일을 저장해야 하는 메모리기반에 비해 메모리 사용측면에서 더 효율적인 장점이 있다[25].

Ⅲ. 연구논문 추천시스템 구축 방안

NDSL은 국내 학계, 연구계, 산업계의 모든 연구자들을 위한 해외 학술저널 및 프로시딩 정보제공 포털로서 2008년 6월 현재 6만 3천여 종의 학술저널과 19만 8천여종의 프로시딩을 서비스 하고 있다[11]. 본장에서는 기존 연구결과와 연계하여 NDSL 연구논문 추천시스템 개발 방안을 모색해 본다.

1. NDSL 개인화 서비스 현황

NDSL에서는 이용자 편의를 위해 개인화서비스를 제공하고 있는데 NDSL에서 서비스하는 정보(논문, 특허, 연구보고서, 동향보고서 등) 중에서 관심 있는 정보를 저장하는 즐겨찾기 기능과 이용자가 등록한 주제나 키워드와 일치하는 정보가 유입(생성)시 이용자에게 알려주는 알림서비스가 있다. 또한 이용자가 검색한 결과와 유사한 주제를 가진 논문을 추천해 주는 기능[그림 4]도 제공하고 있다.

그림 4. NDSL에서 제공하는 알리미 서비스

그러나 이들은 개인화 서비스의 가장 기초적인 단계인 알림서비스[그림 4]에 해당하며 개인이 선호하는 정보와 관련된 프로파일을 명시적으로 입력받아서 추천하는 내용 기반의 필터링 기법이다. NDSL에서는 검색결과와 유사한 정보를 추천해 주는 서비스도 제공하고 있는데 모든 이용자에게 동일한 결과물을 제공한다는 점에서 개인화되지 못한 서비스이다.

2. 이용자 기반과 인용 기반 협업 필터링

개인화된 추천시스템의 선행연구결과에서 NDSL에 응용 가능한 것을 찾기 위해서는 NDSL에서 활용 가능한 자원을 먼저 살펴보아야 한다. NDSL에서는 이용자가 로그인하고 정보서비스를 사용을 하였을 때 이용자 ID, 클릭논문, 다운로드한 논문의 정보 등 시스템 로그파일에 저장된 정보를 추천시스템에 활용할 수 있을 것이다. 그런데 연구논문이 다른 아이템과 차별화 되는 것은 인용이다. 인용은 논문간의 네트워크를 구성하기 때문에 연구논문 추천에 이용될 수 있다. 인용을 추천에 이용하는 경우는 인용과 관련하여 직접적으로 인용받는 참조문헌을 추천하거나 동시인용법[13]이나 서지결합법[2]을 사용한다[1][28]. 이 경우 인용을 하는 논문은 협업필터링에서 이용자에 해당하고 인용을 받는 참조논문은 아이템에 해당한다. 인용을 협업 필터에 사용하는 이유는 협업 필터링에서 발생하는 새로운 시스템과 새로운 아이템 문제를 극복할 수 있기 때문이다. NDSL에서는 전자저널 공동구매컨소시엄(KESLI :

Korean Electronic Site License Initiative)에 포함된 저널에 대해 해외논문정보를 제공하고 있으나 인용정보는 제공하지 않고 있다. 국내에서도 정보원이나 평가도구로 활용하기 위한 인용색인 데이터베이스 구축 사업이 한국과학기술정보연구원, 한국진흥재단, 한국연구재단(구 한국과학재단) 등 기관별로 독립적으로 진행되었으나[4], NDSL은 국내정보에서 인용정보를 제공하지 않는다. 따라서 인용정보를 NDSL 연구논문 추천시스템의 자원으로 활용하는 것은 불가능하다. 향후 인용정보를 제공하는 전자도서관에서는 인용정보를 이용하여 추천시스템을 개발하는 것을 적극적으로 고려해 볼 필요가 있다.

3. 내용기반과 협업 필터링

추천시스템은 이용자가 선호할 것으로 예상되는 아이템을 추천하는 것이다. 선호는 이용자의 아이템을 이용하는 목적에 따라 달라진다[28][29]. 연구논문 정보서비스 이용목적을 연구과정에서 발생한 연구주제나 문제에 대해 넓은 이해를 목적으로 하는지 현재 주제나 문제에 대해 좀 더 깊은 이해를 원하는 지로 나누어 생각해 볼 수 있다. 이 두 가지의 이용목적을 하나의 추천기법으로 달성하는 것은 어렵다[28].

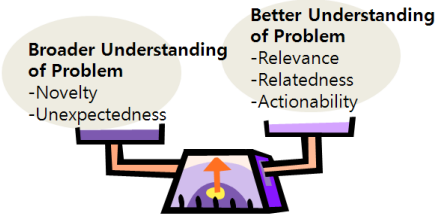

그림 5. NDSL 이용목적과 추천시스템 선정

2002년 수행된 연구[2]에 의하면 NDSL의 이용자의 약 40%가 26세~30세이다(2010년 기준으로 20~30대가 전체의 70%를 차지하고 있다). 이 연령대는 관심 주제에 대해 신규성이 높은 논문보다는 관련성이 높은 논문을 희망할 가능성이 높다. 또한 경험이 많은 교수들에 비해서 경험이 다소 적은 연구자들이 추천시스템에 대한 만족도가 높은 것으로 나타난다[26]. 협업 필터링은 내용기반의 필터링 보다 신규성이 높은 논문을 추천하고, 내용기반의 필터링은 이용자(협업) 필터링에 비해서 주제의 유사성이 높은 아이템을 추천해 주는 경향이 있다[29]. 그러므로 NDSL에서는 이용자 분포를 고려하여 이용자 기반의 필터링 보다는 내용기반의 필터링을 중심으로 시스템을 설계할 필요가 있을 것이다.

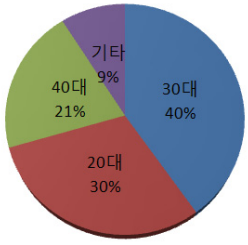

그림 6. NDSL 회원 연령 분포(2010.10 기준)

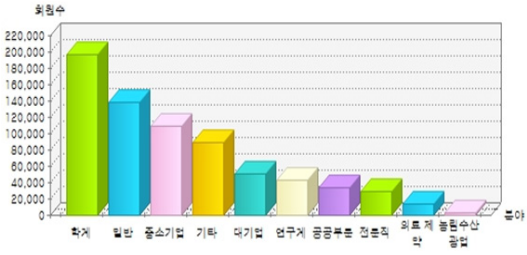

그림 7. NDSL 회원 직업 분포(2010.10 기준)

4. 명시적과 암묵적 방식의 정보수집

NDSL에서 사용하고 있는 명시적 기반의 필터링 방식은 이용자들이 관심 있는 주제나 키워드를 직접 입력해야 추천이 가능하고 이용자들의 관심이 변화할 때마다 이용자들은 번번이 등록한 주제나 키워드를 수정해야 하는 번거로움이 발생한다. 이용자에게 정보검색으로 인해 발생하는 번거로움을 최소화하자는 것이 추천의 주요목적 중에 하나이므로 명시적 방식의 프로파일링 보다는 이용자의 정보 패턴에서 자동으로 프로파일을 추출하는 암묵적 방식이 NDSL 연구논문 추천시스템에 적당할 것이다. 그러나 이용자가 시스템에 최초 가입시 이용자 개인정보입력란에 관심주제나 연구분야를 명시적으로 기입함으로서 새로운 이용자 문제를 방지할 필요가 있다.

5. 이용자 기반과 아이템 기반 협업 필터링

NDSL 이용자는 신규성보다는 유사성이 높은 연구논문 추천을 희망할 가능성이 높지만 신규성이 높은 논문을 찾는 이용자를 서비스에서 배제할 수 없다. 장기적 관점에서 시스템이 신규성이 높은 연구 논문을 추천하는 것은 이용자 폭을 확대하는 데에 기여할 것이다.

신규성이 높은 연구논문을 추천하기 위해서 협업 필터링을 내용 기반의 필터링과 함께 제공할 수 있다. 협업 필터링은 아이템 기반과 이용자 기반 방식으로 구현된다. 아이템 기반 방식은 이용자의 이용패턴을 바탕으로 아이템간의 유사도를 추산하고 이를 추천에 이용하므로 성능과 확장성 측면에서 온라인서비스에 유리하고 추천결과에서도 다소 우수하다고 연구되었다[9]. 이것은 앞서 지적된 협업 필터링이 가지는 확장성 부족의 문제점을 해결할 수 있다는 측면에서 제안된 기법이다. 그런데 일부 연구에서는 아이템 기반의 추천이 이용자 기반보다 이용자 만족도가 낮은 것으로 나타난다[28]. 아이템이나 이용자의 특성과 무관하게 특정 기법이 월등이 낫다는 연구결과가 없는 상황에서는 가장 단순하고 일반적 협업 필터링인 이용자 및 메모리 기반으로 추천 시스템을 구축하고, 이용자를 대상으로 만족도를 분석하면서 조금씩 확장성을 보장할 수 있도록 알고리즘으로 개선해 나가는 것이 NDSL을 비롯한 전자도서관 추천시스템으로 좋을 것으로 판단된다.

6. 혼합형 필터링 방법

내용 기반과 협업 필터링을 동시에 사용하여 유사성과 신규성이 높은 연구논문을 추천하기 위해서는 두 기법을 어떤 방식으로 혼합할 것인가에 대한 결정이 필요하다. 혼합 방법에는 두 알고리즘을 직렬로 연결하는 방법과 두 알고리즘의 결과물을 따로 보여주는 방법, 그리고 두 알고리즘에서 나온 결과물을 합쳐서 하나로 보여주는 병렬적 방법이 있다.

직렬로 두 알고리즘을 연결하는 방식은 이용자의 만족도가 낮다는 연구결과[26]가 있다. 그리고 두 알고리즘에서 나온 결과물을 따로 보여 주면 NDSL의 인터페이스를 복잡하게 하고 이용자들에게 두 추천결과 차이에 대해 혼란을 초래할 것이다. 따라서 두 알고리즘 결과물 중 추정치가 높은 아이템을 골라서 함께 추천을 하는 것이 바람직하다(그림 8). 그런데 이 경우에도 NDSL의 이용자 성향을 고려하여 내용기반 필터링에 좀 더 많은 비중을 두어야 할 것이다(α>0.5).

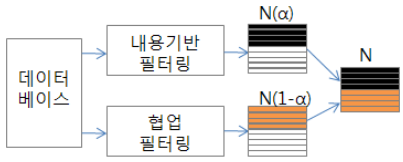

그림 8. 내용기반과 협업 필터링의 혼합

7. 추가 고려사항

NDSL의 알리미서비스로 제공하는 내용기반의 추천 방식을 현재의 명시적 방식에서 암묵적 방식으로 이용자 프로파일 구축방식을 바꿔서 추천을 하면 유사성이 높은 연구논문 추천에 적합할 것이다. 그러나 개인의 프로파일을 이용할 때에도 단순히 이용자가 클릭한 논문의 키워드 벡터 공간만을 이용하는 것보다는 온톨로지 관점에서 접근하여 개인의 프로파일의 특성을 뽑아내거나[27][15], 최근에 클릭한 논문과 그것에 포함된 단어(키워드)에 가중치를 예전에 클릭한 것보다 높게 설정하는 방법(time aging)을 사용하는 것이 성능개선에 효과가 있을 것이다. 프로파일을 이용한 내용기반의 추천에서는 프로파일에서 주요 키워드를 찾는 것이 중요한데 쿼리 기반의 샘플링 방법[17]을 구현 알고리즘으로 고려해 볼 수 있다. 개인의 신상정보를 이용하는 것도 새로 들어온 이용자에게 발생하는 정보 부재의 문제를 해결하는 데에 도움이 된다.

Google의 뉴스 추천시스템 연구[7]에서는 모델방식과 메모리 방식을 결합하는 협업 필터링을 사용하고 Minhash 알고리즘으로 전세계 이용자 간의 유사도를 계산한다. Google이 대규모 이용자들에게 뉴스정보를 효과적으로 추천하는 알고리즘은 향후 NDSL 연구논문 추천시스템 확장성 개선을 위해 참고가 된다. 협업 필터에서 이용 희소성에 의해 발생되는 추천 만족도 저하의 문제는 이용자 신상정보나 개인 프로파일을 이용하여 밀집도가 높은 벡터로 공간을 분할하여 추천 알고리즘을 구현함으로써 완화할 수 있다[10].

Ⅳ. 결론 및 향후연구

본 논문에서는 국내외에서 수행된 추천시스템에 대한 연구결과와 NDSL 이용자 분포를 분석하여 NDSL의 연구논문 추천시스템 구축방안을 모색해 보았다. 현재 NDSL에서 제공하는 알리미서비스를 암묵적 방식으로 바꾸어서 이용자의 프로파일을 구축할 것과 내용기반의 필터링을 중심으로 이용자 및 메모리 기반의 협업필터링을 병행하여 연구논문 추천에서 신규성이 부족한 단점을 보완할 것을 제안하였다. 또한 두 기법을 함께 사용하는 방식과 온톨로지와 분할방식에 의한 필터링을 이용하여 추천 만족도를 높이는 방식에 대해서도 제안하였다.

본 논문은 NDSL의 연구논문 추천시스템 구축 방안을 모색하였으나, 시스템을 구축하여 실제 이용자를 대상으로 만족도를 측정하지 못하였고, 이로 인해 NDSL 데이터와 주이용자의 특성에 맞는 알고리즘을 구체적으로 제안하지 못한 한계를 가진다. 향후연구에서는 최근의 NDSL 이용자의 이용 현황과 목적 및 형태를 분석하고, 추천시스템을 구축하여 이용자를 대상으로 만족도를 평가하여 시스템 구성을 구체화 및 최적화 하고자 한다. 또한 전자도서관이 단순히 보관하고 있는 정보를 인덱스하여 검색기능을 제공하는 수동적 서비스 차원을 탈피하여 연구자들의 문제를 직접적으로 해결해 줄 수 있는 능동적 추천시스템의 구축방안을 제안하고자 한다.

참 고 문 헌

[1] 강남규, 조민희, 권오석, “NDSL 검색 질의어와 기술용어간의 관계에 대한 분석적 연구”, 정보관리연구, 제39권, 제3호, pp.163-177, 2008.

[2] 유사라, “국가과학기술전자도서관 이용자 정보요구와 이용 행태 분석”, 한국문헌정보학회지, 제36권, 제1호, pp.25-40, 2002.

[3] 정영미, 이용구, “필터링 기법을 이용한 도서 추천 시스템 구축”, 정보관리연구, 제33권, 제1호, pp.1-17, 2002.

[4] 최광남, 이재윤, 조현양, “KCI 활용을 위한 지표에 관한 연구”, 정보관리연구, 제37권, 제2호, pp.21-31, 2006.

[5] A. F. Smeaton and J. Callan, “Personalisation and recommender systems in digital libraries”, International Journal on Digital Libraries, Vol.57, No.4, pp.299-308. 2005.

[6] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock, “Methods and Metrics for Cold-start Recommendations”, in Proceedings of 25th annual international ACM SIGIR conference on Research and development in information retrieval, Tampere, Finland, pp.253-260. 2002.

[7] A. S. Das , M. Datar, A. Garg and S. Rajaram, “Google news personalization: scalable online collaborative filtering”, Proceedings of the 16th international conference on World Wide Web, 2007.

[8] B. Rous, “The ACM digital library”, Communications of the ACM, Vol.44, No.5, pp.90-91. 2001.

[9] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, “Item-Based Collaborative Filtering Recommendation Algorithms”, Proc. 10th Int’l WWW Conf., 2001.

[10] F. Gao, C. Xing, X. Du, and S. Wang, “ Personalized Service System Based on Hybrid. Filtering for Digital Library”, Tsinghua Science and Technology, Vol.12, No.1, pp.1-8, 2007.

[11] G. Adomavicius, A. Tuzhilin, “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, Vol.17, No.6, pp.734-749, 2005.

[12] H. Avancini, L. Candela and U. Straccia,

"Recommenders in a personalized, collaborative digital library environment", J. Intell. Inf. Syst. Vol.28, No.3, pp.253-283. 2007.

[13] H. Small, "Co-citation in the scientific literature: a new measure of the relationship between two documents", Journal of the American Society for Information Science, Vol.24, No.4, pp.265 - 269. 1973

[14] http://lib-www.lanl.gov/lww/mylibweb.htm

[15] I. E. Liao, S. C. Liao, K. F. Kao, and I. F. Harn, "A Personal Ontol-ogy Model for Library Recommendation System," in The 9th International Conference on Asian Digital Libraries (ICADL 2006), pp.173-182, 2006.

[16] J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl, "GroupLens: Applying Collaborative Filtering to "Usenet News", Commun ACM, Vol.40, No.3, pp.77-87, 1997.

[17] J. Callan and M. Connell, "Query-based sampling of text databases", ACM Transactions on Information Systems (TOIS), Vol.19, No.2, pp.97-130, 2001.

[18] J. L. Herlocker, J. A. Konstan, A. Borchers and J. Riedl, "An Algorithmic Framework for Performing Collaborative Filtering", in Proc. of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.230-237, 1999.

[19] K. Goldberg, T. Roeder, D. upta, and C. Perkins, "Eigentaste: A Constant Time Collaborative Filtering Algorithm", Inf.Retr., Vol.4, No.2, pp.133-151, 2001.

[20] M. Balabanovic and Y. Shoham. Fab, "Content-Based, Collaborative Recommendation", Communications of the ACM, Vol.40, No.3, pp.66-72. 1997.

[21] M. M. Kessler, "Bibliographic coupling between scientific papers", American Documentation, Vol.14, pp.10-25, 1963.

[22] N. R. Smalheiser, "The Arrowsmith project: 2005 status report", Lecture Notes in Artificial Intelligence. Vol.3735. pp.26-43. 2005.

[23] N. R. Smalheiser, V. I. Torvik, W. Zhou, "Arrowsmith two-node search interface: A tutorial on finding meaningful links between two disparate sets of articles in MEDLINE", Computer Methods and Programs in Biomedicine, Vol.94, No.2, pp.190-197, 2009.

[24] NSF, *N.S.F. Academic Research and Development*. 1999.

[25] R. Jin, L. Si, C. Zhai, and J. Callan, "Collaborative filtering with decoupled models for preferences and ratings", In Proceedings of CIKM 2003, pp.309-106, 2003.

[26] R. Torres, S. McNee, M. Abel, J. Konstan and J. Riedl, "Enhancing digital libraries with TechLens+", In Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries. ACM Press, Tuscon, AZ, USA, pp.228-236. 2004.

[27] S. E. Middleton , N. R. Shadbolt and D. C. De Roure, "Ontological user profiling in recommender systems", ACM Transactions on Information Systems (TOIS), Vol.22, No.1, pp.54-88, 2004.

[28] S. M. McNee, I. Albert, D. Cosley, P. Gopalkrishnan, S. K. Lam, A. Rashid, J. A. Konstan, and J. Riedl, "On the recommending of citations for research papers", In Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work, CSCW '02. ACM, New York, NY, pp.116-125, 2002.

[29] S. M. M.cNee, N. Kapoor, and J. A. Konstan, "on' Look Stupid: Avoiding Pitfalls when Recommending Research Papers", In

Proceedings of the 2006 ACM Conference on Computer Supported Cooperative Work (CSCW 2006), Banff, Canada, 2006.

[30] U. Shardanand and P. Maes, "Social Information Filtering: Algorithms for Automating Word of Mouth", in Proc. of the SIGCHI Conference on Human Factors in Computing Systems, pp.210-217, 1995.

저 자 소 개

여 운 동(Woon-Dong Yeo) 정회원

- 2002년 2월 : 경북대학교 전장공학과(공학석사)
- 2009년 8월 : 고려대학교 컴퓨터학과(박사수료)
- 2002년 4월 ~ 현재 : 한국과학기술정보연구원 선임연구원

<관심분야> : 데이터마이닝, 과학기술 계량정보분석

박 현 우(Hyun-Woo Park) 종신회원

- 1986년 2월 : 홍익대학교 대학원 졸업(경영학석사)
- 1991년 2월 : 홍익대학교 대학원 졸업(경영학박사)
- 2008년 2월 : 고려대학교 대학원 졸업(이학박사)
- 1991년 ~ 1999년 : 산업기술정보원 부연구위원
- 1995년 ~ 1997년 : 미국 San Francisco 주립대 Visiting Scholar
- 2007년 ~ 2008년 : 미국 University of California (Santa Cruz) Research Fellow
- 2000년 ~ 현재 : 한국과학기술정보연구원 책임연구원

<관심분야> : 기술정보 콘텐츠, 가치평가

권 영 일(Young-IL Kwon) 정회원

- 1986년 2월 : 성균관대학교 대학원 졸업(공학석사)
- 2001년 8월 : 성균관대학교 대학원 졸업(공학박사)
- 1991년 ~ 2000년 : 산업기술정보원(KINITI) 책임연구원
- 2001년 ~ 현재 : 한국과학기술정보연구원(KISTI) 책임연구원

<관심분야> : 계량정보분석, 유망기술 발굴, 텍스트마이닝

박 영 욱(Young-Wook Park) 정회원

- 2002년 2월 : 포스텍 전자전기공학과 졸업(공학석사)
- 2002년 2월 ~ 2005년 6월 : 삼성전자 선임연구원
- 2005년 6월 ~ 현재 : 한국과학기술정보연구원 선임연구원

<관심분야> : 과학기술정보의 체계적 제공, 유망기술 발굴