http://dx.doi.org/10.15813/kmr.2017.18.4.004

온라인 구매 행태를 고려한 토픽 모델링 기반 도서 추천[1]

Topic Modeling-based Book Recommendations Considering Online Purchase Behavior

정 영 진 (Jung, Youngjin) 국민대학교 데이터사이언스학과[2]
조 윤 호 (Cho, Yoonho) 국민대학교 경영학부[3]

ABSTRACT

Thanks to the development of social media, general users become information and knowledge providers. But customers also feel difficulty to decide their purchases due to numerous information. Although recommender systems are trying to solve these information/knowledge overload problem, it may be asked whether they can honestly reflect customers' preferences. Especially, customers in book market consider contents of a book, recency, and price when they make a purchase. Therefore, in this study, we propose a methodology which can reflect these characteristics based on topic modeling and provide proper recommendations to customers in book market. Through experiments, our methodology shows higher performance than traditional collaborative filtering systems. Therefore, we expect that our book recommender system contributes the development of recommender systems studies and positively affect the customer satisfaction and management.

Keywords: Topic Modeling; Text Mining; Big data Analysis; Collaborative Filtering; Book Recommender System

1. 서론

정보기술의 발전과 인터넷의 대중화로 온라인 쇼핑 시장의 규모는 지속적으로 증가하고 있으며 인터넷 및 스마트 기술을 활용한 다양한 서비스와 콘텐츠가 새롭게 등장하고 있다. 과거에는 전문 지식을 가진 전문가들이 인터넷 상의 콘텐츠를 생성하는 주체였지만, 최근에는 블로그나 SNS 등의 소셜 미디어가 활성화되면서 일반 사용자가 생성하거나 공유하는 지식의 종류와 양이 폭발적으로 증가하고 있다. 소비자의 구매 의사결정 관점에서 이러한 빅데이터 시대

[1] 논문접수일: 2017년 11월 13일; 2차 수정: 2017년 12월 1일; 게재 확정일: 2017년 12월 4일
[2] 제 1저자
[3] 교신저자

의 다양한 플랫폼의 등장과 정보 공유의 활성화는 의사결정에 참고할 수 있는 정보의 획득이 용이해짐을 의미한다. 하지만 너무 많은 정보의 제공은 오히려 소비자들이 느끼는 주관적 탐색 비용을 증가시켜 의사 결정 과정을 어렵게 만들고 있을 뿐만 아니라 향후 인터넷을 통한 서비스의 지속 사용 의도에도 부정적인 영향을 끼치고 있다 (Haucap and Heimeshoff 2014). 의사결정에 참고할 수 있는 지식 및 정보가 오히려 방대하여 생기는 이러한 정보 및 지식 과부하(information and knowledge overload)를 해결하여 소비자의 만족도를 높이기 위한 대표적인 시도가 추천 시스템이다.

추천 시스템은 사용자의 행동 데이터 및 관련 데이터를 분석하여 추천을 요청한 사용자에게 추천 대상 아이템 목록을 제공하거나 추천 대상 아이템에 대한 해당 사용자의 평가 점수를 예측하는 시스템이다. 즉, 추천 시스템은 데이터 분석 기술 기반의 지식 여과(knowledge filtering) 시스템으로 사용자의 구매 기록 등의 데이터를 기반으로 해당 사용자의 선호를 추론하여 구매를 희망하는 아이템을 쉽게 찾을 수 있도록 지원한다 (Sarwar et al. 2001). 추천 시스템을 통해 소비자는 정보 및 지식 탐색 비용을 낮출 수 있으며 기업은 사용자 관리나 매출 향상 측면에서 다른 기업들에 비해 경쟁력을 갖출 수 있다. 특히 사용자의 선호도를 고려한 적절한 아이템의 추천은 불필요한 정보를 제공하지 않아 자원의 낭비를 줄이고, 사용자의 만족도를 높일 수 있다 (Zhang et al. 2013). 그에 따라 Amazon, Netflix 등의 많은 기업들이 이미 추천 시스템을 자사의 서비스에 적용하여, 데이터를 기반으로 사용자의 needs를 신속히 파악하고 소비자가 선호할 가능성이 높은 콘텐츠 및 아이템을 추천하여 소비자의 구매 의사결정을 지원하고 있다 (현윤진 등 2013; Brynjolfsson et al. 2011; Li et al. 2014).

기존의 추천 시스템 관련 연구에서 가장 성공적으로 평가 받는 기법은 협업 필터링(Collaborative Filtering)으로 웹 페이지, 영화, 신문기사 추천 등 다양한 분야에서 적용되고 있다 (Li et al. 2014; Merve and Arslan 2009; Ricci et al., 2011). 협업 필터링은 추천 대상자의 아이템에 대한 선호를 추론하기 위해 타인의 구매나 평가 정보를 이용하여, 즉 이용자 간 지식의 공유를 통해 추천 서비스를 제공한다. 사용자의 평가나 구매기록에 기반하여 선호를 추론함에 따라 사용자나 아이템에 대한 추가 정보가 존재하지 않아도 추천이 가능하다는 장점이 있지만, 이러한 정보가 목표 고객의 개인화된 선호도를 적절하게 반영할 수 있는지에 대한 의문이 지속적으로 제기되어 왔다 (Zhang et al. 2013). 따라서 최근에는 구매 이력과 동시에 웹 상에서 얻을 수 있는 클릭 스트림 등을 활용하거나 가격 등과 같은 아이템에 대한 정보, 사용자의 인구 통계학 정보 등을 통해 추천 시스템의 성과를 높이려는 시도가 있었다. 또한 사용자 커뮤니티 등의 발전으로 사용자의 이용 후기 공유가 활발해짐에 따라 공유로 생성되는 빅데이터를 활용하여 추천 서비스를 제공하려는 시도도 이루어 지고 있다 (최준연 등 2013). 하지만 이러한 다양한 시도에도 불구하고 도서 추천을 위한 효과적인 시스템을 구축하기 위해서는 소비자의 도서 구매에 영향을 끼치는 요소를 반영할 필요가 있다.

온라인 상에서 도서를 구매하는 사용자는 가장 먼저 해당 책이 어떤 내용을 담고 있는 지 파악한다. 이를 위해 구매하고자 하는 도서의 책 소개 내용이나, 목차 등을 우선적으로 살펴본다. 이는 '국내 소설' 등과 같이 단순한 도서 분류를 참고로 하여 내용을 파악하는 것이 아니라 본인이 관심을 가지는 주제가 책 소개나 목차에 드러나 있는지를 살펴본다는 것을 의미한다. 하지만 기존의 추천 시스템 관련 연구에서는 도서의 책 소개나 목차 등의 고려가 이루어지지 않았다. 다음으로는 다른 상품의 구매 의사결정과 마찬가지로 도서의 최신성을 고려하게 된다. 일반적인 상품은 시간이 지남에 따라 소비자의 구매가 떨어짐에 따라 시간에 따라 추

천될 확률을 낮추는 등 시간의 흐름을 반영한 추천 시스템에 대한 연구는 그간 진행되어 왔다. 하지만 도서의 경우 그 양상이 개인별로 차이가 있을 수 있다. 예를 들어 특정 소비자들은 주제가 본인의 선호에 맞거나 관심이 가는 경우에는 출판 시기에 대한 고려는 하지 않는다. 이는 최신성에 대한 고려가 일괄적으로 적용되는 것이 아니라 개인별 최신성에 대한 선호 정도가 반영되어야 함을 의미한다. 마지막으로 가격 조건에 대한 고려가 필요하다. 특히 온라인에서 도서를 구매하는 경우 책의 세부 내용을 심도 있게 파악할 수 없음에 따라 구매 실패에 대한 비용이 높다. 하지만 가격은 소비자의 경제력과 밀접한 관련이 있어 개인별로 비용에 대한 인식이 다름에 따라 개인별로 가격 수용도가 다르다. 따라서 도서 추천 시스템은 이러한 개인별 가격 수용도를 반영하여 추천 서비스를 제공해야 한다.

본 연구에서는 사용자가 선호하는 주제, 최신성 선호 및 가격 수용도를 반영하여 도서 추천의 성능을 높일 수 있는 새로운 추천 시스템을 제안한다. 이를 위하여 본 연구에서는 먼저 고객 구매기록에 나타난 도서의 소개와 목차에 대해 토픽 모델링을 진행한다. 도서 등과 같은 텍스트 데이터는 대표적인 비정형 데이터로 이로부터 주제를 추출하는 텍스트 마이닝 기법이 토픽 모델링이다. 토픽 모델링에는 다양한 방법이 있지만 대표적으로 가장 많이 사용되는 방법이 LDA(Latent Dirichlet Allocation) 토픽 모델링이다 (박상현 등 2017; 양승준 등 2016; 차윤정 등 2015). 따라서 본 연구에서 제안하는 추천 시스템은 LDA 토픽 모델링을 기반으로 주제를 추출하고 사용자의 선호를 추론한다. 또한 사용자 선호도를 기반으로 유사도를 계산하여 추천 후보 리스트를 생성한 뒤 사용자의 도서 출판의 최신성 선호 여부와 가격 수용도를 기준으로 필터링하여 최종 추천 리스트를 제공한다. 본 연구에서 제안하는 도서 추천 시스템은 새로운 기법의 적용으로 전반적인 추천 시스템 성능 향상에 기여할 것으로 기대되며 사용자의 만족도 향상과 관리에도 긍정적인 영향을 미칠 것으로 기대된다.

2. 이론적 배경

2.1 추천 시스템

추천 시스템은 개인화된 정보 탐색으로 개인의 의사결정을 돕는 도구로 방대한 데이터로부터 사용자가 선호하는 정보만을 선택적으로 제공하는 시스템이다 (김경재·안현철 2009; Sarwar et al. 2001). 최근 추천 시스템은 아마존과 넷플릭스의 아이템, 페이스북의 친구 추천 등 온라인 서비스에 적용되면서 지속적으로 그 응용 영역이 확대되어 가고 있으며 빅데이터에 대한 관심과 함께 그 중요성이 부각되고 있다 (김민정·조윤호 2015). 추천 시스템은 그 원리에 따라 크게 내용기반 필터링 (Content-based Filtering; CBF)과 협업 필터링 (Collaborative Filtering; CF)로 구분할 수 있다.

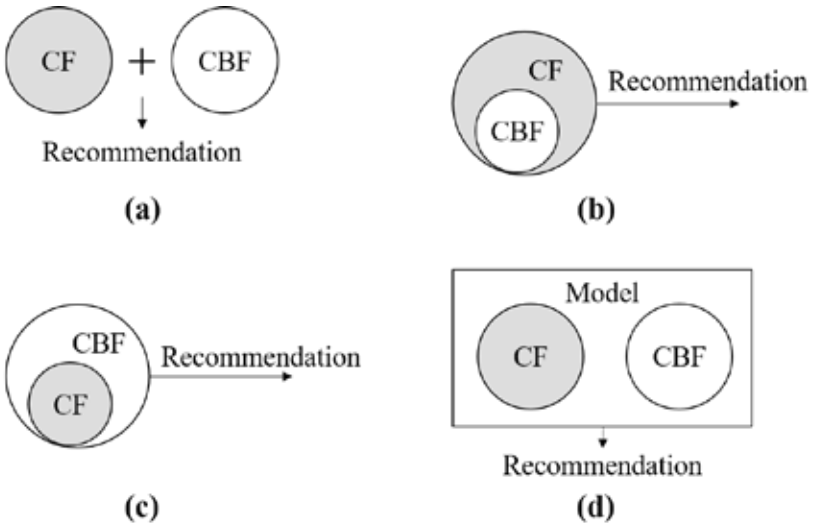
내용기반 필터링은 사용자가 소비하는 아이템의 속성 정보를 분석해 아이템과 사용자 선호도간 유사성을 토대로 추천을 하는 방법이다(Pazzani 2007; Wu and Chen 2000). 많은 내용기반 필터링 관련 연구들은 아이템의 속성을 정의하기 위해 배우, 감독, 장르 등과 같이 속성 값이 명확히 정의되어 있는 정형 데이터를 사용해왔다. 최근에는 이미지, 소리, 텍스트와 같은 비정형 데이터에 대한 연구가 활발히 이루어 지면서 문서, 웹사이트 등의 텍스트로 이루어진 콘텐츠에 대해 속성을 정의하고 추천 방법론을 설계하기 위한 연구들이 진행되고 있다 (손지은 등 2015).

협업 필터링은 사용자들 간의 구매 패턴 유사성을 분석하여 아이템의 선호도를 추론하는 방법으로, 기존의 사용자 구매 정보를 분석하여 해당 사용자와 비슷한 성향의 사용자들이 기존에 선호했던 아이템을 추천하는 방식이다 (Herlocker et al. 2004). 이는 col-

lective knowledge의 일종인 구전 효과의 원리를 이용한 것으로서, 추천 대상이 되는 고객과 유사한 구매 패턴을 가지는 고객들을 파악하고 이들이 선호하는 아이템 중에 추천 대상이 되는 고객이 구매하지 않은 아이템을 추천하는 방식이다. 협업 필터링은 다시 유사도를 계산하는 대상에 따라 사용자 기반 협업 필터링(User-based CF; UBCF)과 아이템 기반 협업 필터링(Item-based CF; IBCF)로 구분된다. UBCF는 목표 사용자와 유사한 구매 이력을 보이는 이웃 집합(Neighbor Set)을 파악하고, 목표 사용자에게 이 그룹에 속한 사용자들이 공통적으로 높게 평가한 아이템을 추천한다. 하지만 각 사용자들과 다른 모든 사용자들 간의 유사도를 계산해야 함에 따라 사용자 수에 따라 연산에 필요한 시간과 비용이 증가하는 단점이 있다 (임일 2015). IBCF는 사용자들의 평가 패턴을 바탕으로 아이템들 간의 유사도를 계산한 후, 사용자가 어떤 아이템을 구매하거나 높게 평가하면 그 아이템과 유사한 아이템을 추천하는 방식이다. 특히, IBCF는 아이템 간의 유사도를 계산하는 방식으로 사용자들 간의 유사도를 계산하는 UBCF에 비해 연산 속도가 빠른 것으로 알려져 있다.

협업 필터링은 내용 기반 필터링에 비해 상대적으로 더 우수한 추천 정확도를 보이고 적용의 용이성으로 개인의 맞춤형 서비스가 필요한 광고나 웹페이지 구성 등 다양한 분야에 적용되고 있다 (김경재·안현철 2009). 그러나 협업 필터링은 구매 기록이 많은 경우에 연산량이 기하급수적으로 증가하여 시스템의 확장성(Scalability) 문제가 발생되며 사용자의 선호도가 입력되지 않은 아이템의 개수가 많을 경우 이웃 집합을 탐색하는 데 한계가 발생되는 데이터의 희박성(Sparsity) 문제 등을 가지고 있다 (김경재·안현철 2009). 이러한 한계를 해결하기 위해 최근에는 협업 필터링과 내용기반 필터링의 장점을 극대화하면서 단점을 보완하여 사용자가 필요로 하는 아이템을 효율적으로 찾을 수 있는 하이브리드(Hybrid) 추천 시스템에 대한 연구가 활발히 이루어지고 있다 (손지은 등 2015; Barragáns-Martínez et al., 2010; Burke, 2002; Cho and Kim 2004)

특히 Adomavicius and Tuzhlin (2005)은 CF와 CBF를 결합하는 하이브리드 기법을 <그림 1>과 같이 네 가지 형태로 분류하였다.



Source : Recommender Systems Survey (Bobadilla et al. 2013)

<그림 1> 하이브리드 기법의 분류

먼저, (a)는 CF와 CBF의 독립적인 추천 결과를 결합하는 방식이다 (Burke 2002). (b)는 사용자의 평가 점수 대신에 내용기반 사용자 프로파일을 이용하는 것으로 CBF를 통해 사용자의 선호를 점수화하는 방식이다. (c)는 반대로 협업 필터링으로부터 얻을 수 있는 정보를 CBF에 융합하여 적용하는 것으로 일반적으로 협업 필터링의 추천 결과를 아이템의 속성에 따라 필터링한다. 마지막으로, (d)는 CF와 CBF를 동시에 고려하여 단일 모델을 구축하는 방식을 의미한다. 본 논문에서는 도서 추천 시스템의 성과를 높이기 위해 도서의 목차와 책 소개에 해당되는 내용 기반 정보를 CF에 적용하는 (b) 형태의 하이브리드 기법과 동시에 가격과 최신성에 대한 민감도를 최종 추천리스트에 반영하는 (c)형태의 하이브리드 기법을 사용하였다.

2.2 토픽 모델링

디지털 도서관의 책이나 연구 논문, 온라인 뉴스, 이메일, SNS 등 온라인 상에 다양한 텍스트 데이터들은 그 형태나 크기가 일정하지 않은 대표적인 비정형 데이터이다. 비정형 데이터의 크기는 SNS 등의 발전으로 그 양이 급증하고 있지만 정형 데이터의 분석과는 다른 접근이 필요하다. 그에 따라 비정형 데이터 중 텍스트 데이터를 정형화하고 지식을 추출하는 텍스트 마이닝 기법에 대한 관심이 높아지고 있다.

텍스트 마이닝(Text Mining)은 자연어로 구성된 비정형 텍스트 데이터에서 패턴 또는 관계를 추출하여 의미와 가치가 있는 정보를 찾아내는 데이터 마이닝 기법이다. 이는 사람들이 사용하는 언어를 컴퓨터가 이해할 수 있도록 하는 자연어 처리(Natural Language Processing)에 기반을 두고 있으며 (현윤진 등 2013), 특정 상품이나 서비스에 대한 선호도 및 여론의 방향 등을 파악하는 데 활용되고 있다. 텍스트 데이터는 구조화되어 있지 않아 그 분석에 어려움이 많아 문서별 사용된 용어의 빈도수를 파악하여 문서의 주제 및 특정 단어를 손쉽게 요약할 수 있는 벡터 공간 모델(Vector space model) 기반의 TF-IDF(Term Frequency-Inverse Document Frequency) 기법 등을 사용한다 (Balabanović and Shoham 1997; Salton and Buckley 1988). TF-IDF는 여러 문서에서 자주 출현하는 일반 단어는 가중치를 낮추고 특정 문서에만 출현하는 비 일반 단어의 가중치는 높게 부여하는 계산 방식으로 각 문서는 용어 수만큼의 차원과 TF-IDF를 값으로 갖는 벡터로 표현된다. 하지만 문서 군 내에 존재하는 용어의 수가 방대함에 따라 SVD(Singular Value Decomposition) 등의 차원 축소 기법을 사용한다. 이러한 과정을 통해 비정형 문서의 구조화가 완료되면 정형 데이터와 함께 텍스트 데이터에 대한 군집화, 예측 등의 데이터 마이닝 작업이 가능해 진다.

토픽 모델링(Topic Modeling)은 구조화된 텍스트 데이터에서 각 문서의 주제를 탐색하는 방법으로 최근 텍스트 마이닝에서 대표적으로 사용되는 기법이다. 토픽 모델에서 각 문서는 각자 다른 비율을 가진 여러 개의 토픽으로 구성된다고 가정한다. 이 가정을 기반으로 제시된 토픽 모델은 Hofmann(1999)이 제시한 주성분 분석(Principal Component Analysis; PCA)기반의 Probabilistic Latent Semantic Analysis(pLSA) 등 여러 기법들이 있지만 최근 가장 널리 사용되는 방법은 베이지안 Mixture model 기반의 Latent Dirichlet Allocation (LDA) 기법이다 (Blei and Lafferty 2009). LDA기법은 각 문서가 다루는 주제가 비슷한 문서들을 군집화하여 토픽을 모델링하는 기법으로 주어진 문서가 가지고 있는 잠재적인 주제들과 한 단어가 주제에 포함될 가능성을 디리클레 분포에 기반하여 추정하는 기법이다. 즉, LDA는 용어의 자질을 고려하여 의미적 일관성 있는 주제를 생성하 기 위해 각 문서별 용어들의 분포를 분석하여 주제 군집에의 포함 여부를 판단한다(López et al. 2015; Sheydaei et al. 2015; Xin

2013). 토픽 모델링 기법은 유사한 주제를 갖는 문서들을 묶어서 하나의 토픽으로 구성한다는 점에서 군집화 기법과 유사한 특성을 가지지만, 하나의 문서가 다수의 토픽에 대응된다는 점에서 기존의 군집화 기법과 차별성을 갖는다. 그에 따라 신문이나 논문에서 주제를 추출하여 경향을 분석하거나 다양한 학문 분야에서 핵심 이슈 문서 도출 및 응용하는 데 LDA 기법이 널리 활용되고 있다 (Griffiths and Steyvers 2004; Mimno et al 2008; Song and Kim 2013;).

최근에는 이러한 토픽 모델링 기법의 발전을 추천 시스템에 활용하려는 시도도 이루어 지고 있다. 먼저, Wilson et al. (2014)는 아이템의 장르나, 플롯 등의 아이템 카탈로그의 문구를 LDA기반의 토픽 모델링 기법으로 산출한 결과의 유사도와 사용자 평점의 유사도를 사용하였다. 계산된 유사도 결과를 곱한 값으로 이웃 집합을 탐색하고 추천하는 방법을 제안한 결과 MovieLens와 넷플릭스 데이터 모두에서 기존의 협업 필터링 기법보다 높은 성과를 보였다. 또한 Zhou and Wu (2016)는 사용자들의 행동이 다른 사람들의 행동이나 의견에 영향을 받는다는 점에서 착안하여 LDA 결과 값에 사용자 관심도, 관심 항목, 등급 정보를 결합하여 유사도를 계산하여 추천 과정을 진행하는 RLDA 모델을 제안하였으며 MovieLens 데이터에 높은 성과를 나타냈다. 많은 연구에서 토픽 모델링을 협업 필터링 기법과 결합하여 사용한 결과 높은 성과를 나타냄에 따라 본 연구에서는 책 소개와 목차 내용에서 주제를 추출하여 아이템의 특성을 파악하기 위해 LDA 기반의 토픽 모델링 기법을 적용하였으며 이를 통해 도출된 결과를 협업 필터링 기법과 결합하여 도서 추천 시스템의 성과를 높이고자 한다. 특히 Bakos (1997)과 Peterson et al., (1997)의 연구에 따르면 온라인에서 상품 정보는 사용자의 탐색 비용을 줄일 수 있어 사용자의 만족도를 높인다. 또한 Park and Kim (2003)의 연구에서는 상품 정보의 품질이 구매 행동에 영향을 미친다고 하였다. 따라서, 사용자가 구매를 한 이유가 도서의 상품 정보인 책 내용과 목차에서 주제를 추출하여 이를 추천 시스템에 반영하게 되면 추천 시스템의 성과가 높아질 수 있으리라 기대된다.

2.3 온라인 사용자 구매 행태

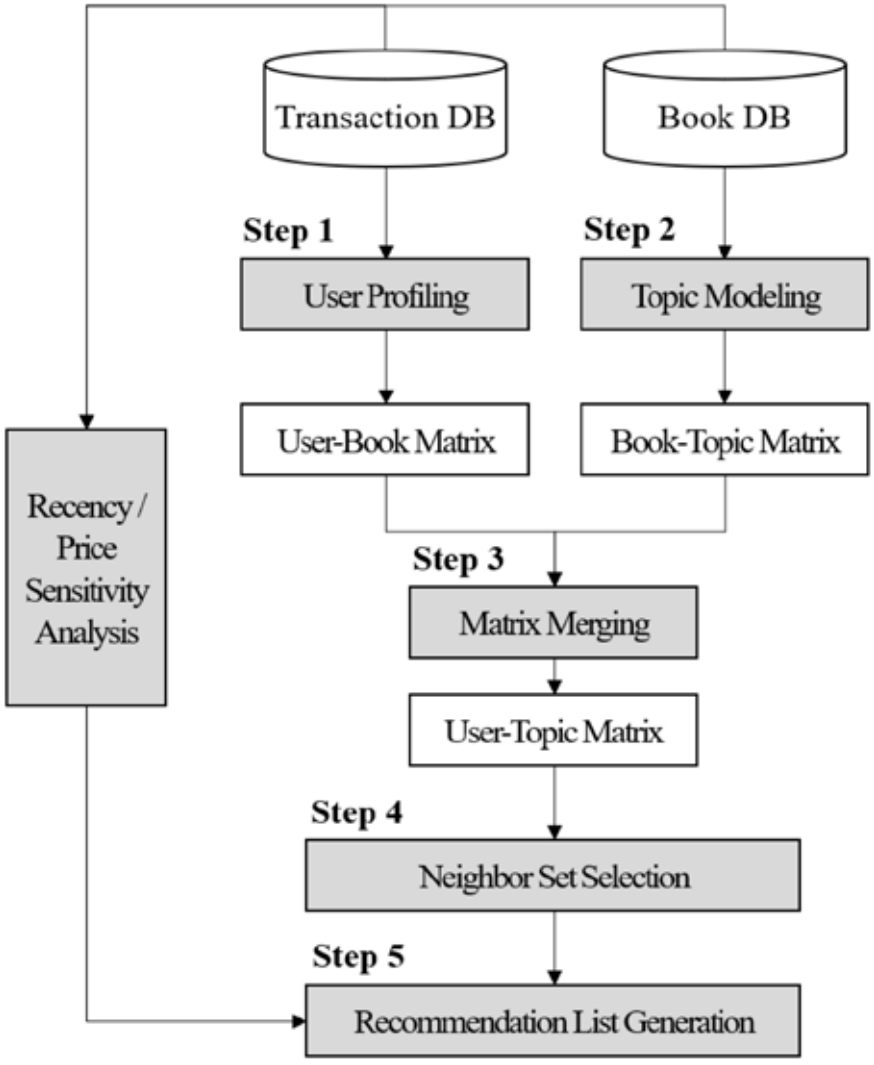
사용자의 구매 행태는 연령과 성별, 직업과 같은 인구 통계학적 특성과 개인의 흥미 영역 등에 따라 많은 행태가 존재하지만 본 연구에서는 도서 추천 시스템의 범위에 따라 도서 구매 행태에 큰 영향을 미치는 가격과 최신성에 따른 온라인 사용자의 구매행태에 관련된 연구를 제시한다. 먼저, 가격은 사용자의 선호도와 구매를 결정하는 데 있어 중요도가 높은 요소로 대부분의 연구는 가격 비교 및 마켓 상품의 가격 결정에 초점을 맞추고 있다 (전지은·이충권 2014; Bacos, 1997). 예를 들어 이세윤 등(2006)의 연구에서는 개인이 과거의 경험에 따라 가격 수용폭을 가지고 있으며 시간의 변화에 따라 새로운 준거 가격을 설정한다는 점에 기반하여, 인터넷 영화관을 대상으로 전송 품질, 가격 수용폭을 벗어난 경우의 가격 민감도, 6개월 단위의 시간에 따른 최신성, 부가 콘텐츠를 활용하여 기업의 서비스 가격 책정 전략을 위한 가격 반응 함수를 분석하였다. 연구 결과 가격의 수용폭은 이전에 구매한 최소값과 최대값 사이의 구간에서 나타나며 이 영역을 벗어난 경우 가격에 민감해지는 경향을 파악하였다. 또한 전지은·이충권 (2014)은 시간이 흐름에 따라 사용자들의 가격에 대한 민감성이 높아질 것이라 가정하여 온라인 마켓에서 일, 주 간격의 시간 흐름에 따른 사용자의 패턴을 분석하여 판매자의 가격 조정 전략을 제안하였다. 앞선 연구들은 판매자의 상품 가격 책정 전략을 위한 연구가 대부분으로 본 연구에서는 이세윤 등(2006)의 연구에 따라 고객이 이전에 구매한 도서의 최소값과 최대값의 구간을 가격 수용도의 기준으로 하여 추천에 적용하고자 한다.

가격과 더불어 시간의 흐름에 따른 추천 아이템의 속성과 사용자의 선호 변화에 대한 고려도 필요하다. 예를 들어 계절이나 일시적 유행에 따라 구매가 달라지고 일반적으로 최근에 등장한 아이템에 대한 구매가 더 많으며, 일정 시점 이후에는 거의 소비가 이루어지지 않기도 한다. 이러한 아이템의 최신성을 추천 시스템에 반영하고자 하는 시도는 꾸준히 이루어져 왔다. 예를 들어, Billsus and Pazzani (1999)는 목표 고객이 관심을 가질 가능성이 높은 뉴스를 추천하기 위해 짧은 기간 동안의 뉴스 구독 기록을 기반으로 하는 추천 모델과 오랜 기간 동안의 뉴스 구독 기록을 기반으로 추천하는 모델을 제안하였다. 그 결과 2개의 모델을 독립적으로 사용하는 것보다 적절한 결합이 더 높은 성과를 보였다. 또한, Sugiyama et al. (2004)는 목표 고객의 하루 동안의 검색 기록에 기반하여 용어 가중치를 계산하고 이를 프로파일화하여 목표 고객의 정보 요구에 적합한 검색 결과를 필터링할 수 있는 시스템을 제안하였다. 그 결과 시간에 따라 변화하는 고객의 정보 선호가 지속적인 선호를 반영한 것보다 더 높은 성과를 나타내었다. Ding and Orlowska (2006)의 연구에서는 목표 고객의 변화된 취향을 반영하기 위해 IBCF에 목표 고객이 상품에 대해 평가한 실제 평가값과 고객의 변화된 취향을 반영한 예측 평가값을 결합하는 추천 방법론을 제안하였다. 마지막으로 Moon et al. (2017)은 고객의 선호가 시간의 흐름에 따라 변화한다는 점에 주목하여 온라인과 오프라인 데이터를 이용하여 시장 환경에 따라 데이터의 최신성이 추천 성과에 미치는 영향을 비교하였다. 분석 결과 온라인 환경이 오프라인 환경에 비해 데이터의 최신성에 더 많은 영향을 받는 것을 확인하였다. 기존의 연구들은 사용자의 시간에 따른 선호도 변화를 추천 시스템에 반영하고 그 효과를 확인하였다. 하지만 Billsus and Pazzani (1999)의 연구 결과에서 나타난 것처럼 최신성에 대한 반응은 개인마다 차이가 있다. 특히 도서의 경우 베스트 셀러를 선호하는 사람이 있는 반면 고전을 선호하는 사람이 있을 정도로 그 반응에 차이가 많다. 따라서 본 연구에서는 도서가 구매된 일자를 기준으로 출판된 일자와의 차이를 계산하여 이를 최신성 측면에서 추천 시스템에 반영하고자 한다.

3. 도서 추천 방법

본 연구에서 제안하는 도서 추천 시스템인 BCBCF (Book Contents-Based Collaborative Filtering)의 전체 프로세스는 <그림 2>와 같다.

먼저, Step 1에서는 고객 구매기록 DB로부터 사용자의 도서 이용 선호를 파악하고 사용자-도서 매트릭스를 생성하여 사용자를 프로파일링한다. Step 2에서는 도서 DB로부터 책 소개와 목차를 수집하여 저장한 코퍼스의 자료로 형태소 구분을 하고 키워드를 추출한 뒤 LDA 토픽 모델링을 이용하여 각 도서의 토픽을 분석한다. 토픽 분석의 결과로 각 도서가 토픽에 속할 확률을 나타내는 도서-토픽 매트릭스를 생성한다. Step 3에서는 앞서 생성한 사용자-도서 매트릭스와 도서-토픽 매트릭스를 결합하여 사용자-토픽 매트릭스를 생성한다. Step 4에서는 사용자-토픽 매트릭스를 기반으로 사용자 간의 유사도를 계산하여 목표 사용자의 이웃 집합을 탐색한다. 마지막으로 Step 5에서는 최종 추천 리스트를 생성한다. 이를 위하여 먼저 이웃 집합 내 사용자들의 구매 기록을 기반하여 목표 고객을 위한 후보 추천 리스트를 생성하고 고객 구매행태 정보를 반영하여 최종 추천 리스트를 생성한다. 이를 위해 목표 사용자가 해당 카테고리 도서의 출판 최신성을 선호하는지 여부에 따라 필터링을 한 뒤 이전에 구매했던 가격범위에 따라 가격 수용도를 반영하여 최종 추천 리스트를 생성한다.



<그림 2> 도서 추천 프로세스

3.1 사용자 프로파일링

사용자 도서 선호 분석 단계에서는 고객 DB의 데이터를 분석하고, 사용자가 도서를 구입했는지를 나타내어 사용자-도서 매트릭스(Matrix)를 도출한다.

본 연구에서는 식(1)에서와 같이 특정 도서에 대한 사용자의 구매가 있으면 1, 없으면 0으로 표시하였다.

$$U_{xi} = \begin{cases} 1 : \text{사용자 } x\text{가 도서 } i\text{를 구매} \\ 0 : \text{사용자 } x\text{가 도서 } i\text{를 구매하지 않음} \end{cases} \quad (1)$$

따라서, 사용자 *x*가 도서 를 구매(*Uxi*=1)했다면 사용자 *x*가 도서 *i*를 선호하고 있음을 가정하고 *Uxi*=0이면 해당 도서에 대해 선호하지 않거나 모른다는 것을 의미한다.

예를 들어 <표 1>과 같은 사용자 도서 구매 내역이 있을 경우 식(1)을 적용하게 되면 <표 2>와 같은 사용자-도서 매트릭스를 도출할 수 있다.

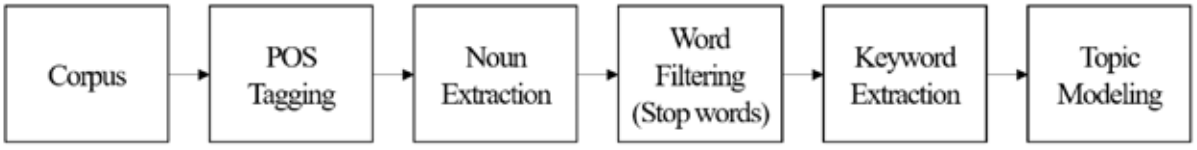
<표 1> 사용자 도서 구매 내역 예제

User	Books	User	Books
U1	{B1, B6}	U6	{B5}
U2	{B2, B3, B5}	U7	{B3, B6}
U3	{B1, B4}	U8	{B4}
U4	{B2, B6}	U9	{B2}
U5	{B1, B4, B7}	U10	{B1, B2, B5, B7}

<표 2> 사용자-도서 매트릭

	B1	B2	B3	B4	B5	B6	B7
U1	1	0	0	0	0	1	0
U2	0	1	1	0	1	0	0
U3	1	0	0	1	0	0	0
U4	0	1	0	0	0	1	0
U5	1	0	0	1	0	0	1
U6	0	0	0	0	1	0	0
U7	0	0	1	0	0	1	0
U8	0	0	0	1	0	0	0
U9	0	1	0	0	0	0	0
U10	1	1	0	0	1	0	1

3.2 토픽 모델링



<그림 3> 토픽 모델링 프로세스

이 단계에서는 <그림 3>과 같이, 먼저 도서 DB의 데이터에서 도서의 책 소개와 목차로 구성된 코퍼스에서 각 단어의 형태소 분석(Part-of-Speech Tagging)을 수행한다. 그 중 명사들을 별도로 추출한 뒤 주제를 분석하기에 적합하지 않은 불용어(stop words)와 '책', '속' 등과 같이 문서에 등장하는 단어 중 빈도수는 많지만 의미가 불분명한 단어들을 제거하여 <표 3>와 같이 각 도서별로 키워드를 추출한다.

다음으로 LDA 알고리즘을 통해 도서내용의 토픽들을 추출하고, 그와 관련된 키워드 및 문서를 추출한다. 각 도서별로 책 소개와 목차에서 추출한 키워드들을 이용하여 LDA 토픽 모델링을 하였을 때 추출된 토픽(Topic words)의 예는 <표 4>에 나타나 있으며, 이를 통해 각 토픽이 어떠한 키워드들로 구성되는지 파악할 수 있다.

<표 5>는 도서별로 각각의 토픽에 연관된 정도를 나타낸 예제이다.

<표 3> 키워드 추출 예제

Book	키워드 (일부)									
B1	엄마	육아	아이	집	경험	과정	낮잠	발	성장	어린이집
B2	생각	방법	이야기	개념	논리	도움말	동안	여행	자연	잘못
B3	쓰기	글쓰기	엽서	마녀	편지	다양	방법	사고력	구미	기사문
B4	여행	중국어	구성	기초	단어	문장	손가락	핸드북	간편	관광
B5	그레그	문제	수영	아이	이야기	일기	가족	갈등	그림	마음
B6	바다	바닷속	나비	눈앞	무시	생물	체험	친구	환상적	고기
B7	거위	깃털	모험	문제	수상	슬픔	작품	장애자	제기	행복

<표 4> 토픽 예제

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
스티커	기업	운동	하나님	역사	작품	경제
0.022217	0.012591	0.03057	0.024749	0.02809	0.013802	0.036291
만화	투자	실험	성경	시대	작가	사회
0.014159	0.011956	0.016683	0.014907	0.015623	0.013684	0.022451
놀이	시장	다이어트	예수	조선	소설	경제학
0.010506	0.009366	0.014363	0.01403	0.012108	0.013342	0.01748
공룡	전략	과학	인류	그림	이야기	주의
0.009961	0.007802	0.01371	0.013575	0.009906	0.012814	0.012933
캐릭터	성공	커피	교회	한국사	사건	자본주의
0.008212	0.006625	0.009924	0.009955	0.008788	0.010107	0.009206
색칠	경영	만들기	인간	미술	사랑	국가
0.008194	0.006316	0.008641	0.009219	0.00839	0.007942	0.009118
활동	미래	건강	기도	문화	사람	정치
0.008161	0.006176	0.007922	0.00809	0.007663	0.006687	0.008147

<표 5> 토픽별 도서 연관 정도

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
B1	0.036056	0.013859	0	0	0	0.016879
B2	0.032857	0.099350	0.012893	0	0	0
B3	0.036897	0	0.016971	0	0	0.066234
B4	0	0	0	0	0	0
B5	0.010151	0	0	0	0.223083	0.356177
B6	0.054069	0	0	0.032298	0	0
B7	0.011591	0	0	0.254385	0.049331	0

<표 6> 도서-토픽 매트릭스

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
B1	1	1	0	0	0	1
B2	1	1	1	0	0	0
B3	1	0	1	0	0	1
B4	0	0	0	0	0	0
B5	1	0	0	0	1	1
B6	1	0	0	1	0	0
B7	1	0	0	1	1	0

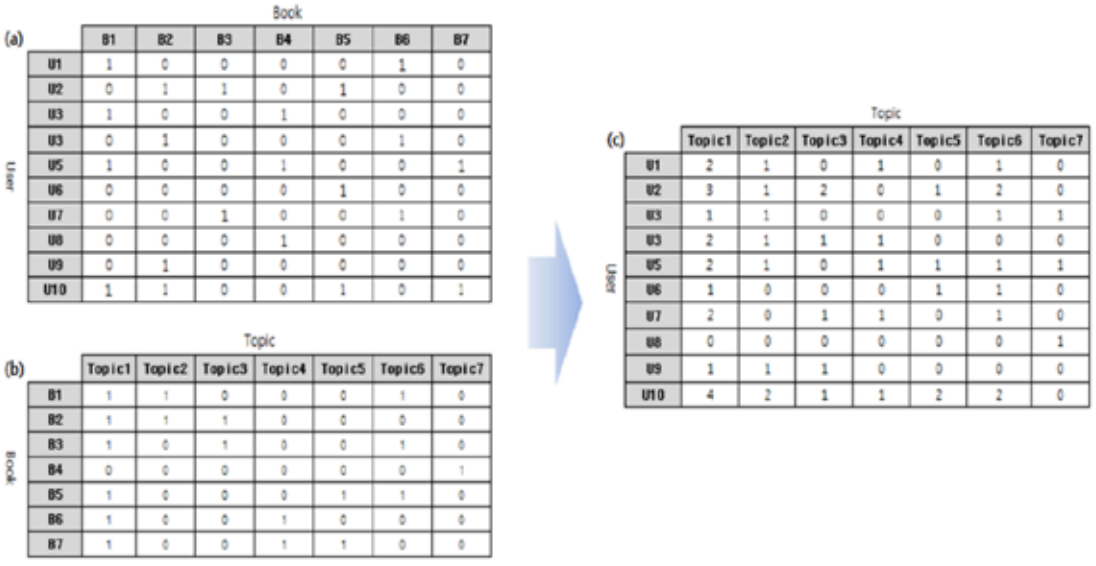
본 논문에서는 현윤진 등(2013)의 연구방법을 적용하여, 식(2)와 같이 도서i와 토픽j의 연관성이 0보다 크다면 도서*i*가 토픽*j*에 해당된다고 파악하여 1로 나타내고, 반대로 연관성이 0이면 도서*i*가 토픽*j*에 해당되지 않는다고 판단하여 0으로 표시하였다. <표 6>은 도서를 해당 토픽에 적용시킨 도서-토픽 매트릭스이다.

$$B_{ij} = \begin{cases} 1 : \text{도서 } i\text{와 토픽 } j\text{의 연관성} > 0 \\ 0 : \text{도서 } i\text{와 토픽 } j\text{의 연관성} = 0 \end{cases} \quad (2)$$

3.3 매트릭스 병합

앞서 도출된 사용자-도서 매트릭스와 도서-토픽 매트릭스를 식(3)과 같이 병합하여 사용자와 토픽 간의 대응 매트릭스를 도출한다. 두 매트릭스의 병합 과정을 도식하면 <그림 4>와 같다.

$$T_{xj} = U_{xi} \cdot B_{ij} \quad (3)$$



출처 : 토픽 분석을 활용한 관심 기반 고객 세분화 방법론 (현윤진 등, 2015)

<그림 4> 매트릭스 병합

<그림 4>와 같이 (a)사용자-도서 매트릭스와 (b)도서-토픽 매트릭스를 곱하여 (c)의 사용자-토픽 매트릭스를 구성한다. 행렬곱 기반의 이 매트릭스를 통해 사용자가 각 토픽에 해당되는 도서의 구입 여부와 사용자가 해당 도서가 가지는 토픽의 주제에 관심(선호)이 있는지를 동시에 파악할 수 있다.

3.4 유사도 분석 및 이웃 집합 구성

다음으로 병합한 사용자-토픽의 매트릭스를 이용하여 사용자간 유사도를 계산하여 이웃 집합을 찾고 대상 사용자가 아직 구매하지 않은 도서에 대한 구매정보를 예측한다. 협업 필터링 기법에서 사용자들 사이의 유사도를 산출하는 작업에는 많은 유사도 계산식이 있지만, 연속 값에 대해서 가장 널리 쓰이는 유사도 측정값으로는 코사인 유사도가 있다. 본 연구에서는 코사인 유사도를 위하여 각 주제를 차원으로 보고 사용자의 구매횟수를 좌표로 본다. 그렇게 되면 각 사용자의 값을 벡터로 해서 두 사용자 간의 벡터의 각도(코사인 값)을 구할 수 있다. 식(4)는 코사인 유사도 기반의 유사도 계산을 나타낸다.

$$S_{x,y} = \frac{\vec{x}\cdot\vec{y}}{\|\vec{x}\|\cdot\|\vec{y}\|} = \frac{\sum_{j\in C} T_{xj}\cdot T_{yj}}{\sqrt{\sum_{j\in C} T_{xj}^2}\sqrt{\sum_{j\in C} T_{yj}^2}} \quad (4)$$

(4)에서 x와 y는 유사도 계산 대상이 되는 두 사용자이고, T_{xj}, T_{yj}는 두 사용자가 각각 선호하는 토픽을 나타낸다. C는 사용자가 연관된 토픽 전체를 나타낸다. 식(4)를 통해 계산된 유사도 값은 <표 7>과 같이 최소 0(완전 불일치)에서 최대 1(완전 일치)의 값을 갖는다.

<표 7>과 같은 코사인 유사도 매트릭스를 기반으로 이웃 집합(Neighbor Set)을 정하는 방법은 크게 두가지 방법이 존재하는데, 하나는 이웃 집합의 크기를 미리 정하고 추천 대상 사용자와 가장 유사도가 높은 N명을 선택하는 Top-N Best Neighbors(또는 Best-N Neighbors)방법이고, 또 다른 하나는 크기 대신 유사도의 기준을 정해 놓고 이 기준을 충족시키는 사용자를 이웃 집합으로 정하는 Thresholding이다. Thresholding 방법은 정해진 기준을 넘는 경우가 불규칙함에 따라 주로 Top-N 방법이 사용된다 (임일 2015). 따라서 본 연구에서는 추천 대상 사용자와 가장 유사도가 높은 순으로 N명의 사용자를 이웃 집합으로 구성한다.

<표 7> 코사인 유사도 매트릭스

	U1	U2	U3	U4	U5	U6
U1	1	0.3834	0.3549	0.2757	0.4934	0.3119
U2	0.3834	1	0.1683	0.4445	0.0755	0.0185
U3	0.3549	0.1683	1	0.2453	0.4956	0.1301
U4	0.2757	0.4445	0.2453	1	0.0825	0.0000
U5	0.4934	0.0755	0.4956	0.0825	1	0.0875
U6	0.3119	0.0185	0.1301	0.0000	0.0875	1

3.5 최신성과 가격을 반영한 추천 리스트 생성

도서는 한번 구매한 경우 재구매를 하지 않고 새로운 제품을 구매하는 특성이 있으므로 이웃 집합의 사용자들이 구매했던 도서 중에서 목표 사용자가 이전에 구매하지 않았던 도서를 선정하고 필터링하여 Top-k개의 도서를 추천한다.

본 연구에서는 최종 추천 도서 리스트를 생성하기에 앞서 사용자의 구매행태 중 최신성과 가격 수용도를 반영하고자 한다. 먼저, 사용자가 도서를 선택할 때 최근에 출간된 도서를 얼마나 선호하는지 반영하기 위하여 도서의 출판일로부터 판매일까지의 기간을 재고기간으로 각 카테고리별 중위수를 구하여 계산하고, 사용자가 구매한 카테고리별 도서의 재고기간이 도서 전체 카테고리의 재고기간보다 짧다면, 사용자는 그 카테고리의 도서에 대해 최신성을 선호한다고 판단한다. 산출한 후보 목록 중 사용자가 최신성을 선호하는 카테고리의 도서라면, 최신성을 갖지 않는 도서는 제외한다. 가격 수용도는 사용자가 이전에 구매한 도서의 가격 범위인 최고값~최저값으로 계산한다. 즉, 최저값보다 적은 가격의 도서와 최고값보다 높은 가격의 도서는 추천 대상에서 제외한다.

도서의 최신성과 가격 수용도를 기준으로 필터링하는 알고리즘은 다음과 같다.

Algorithm RecommendFiltering(Recency and Price)

```
Input:   I=Set of books,   x=User,
         Ix=Book of x(saleDate-publishDate), Ic=Category of book,
         Rc=Recency catogory book of x liked (0 or 1)
         L= Recommend Book List (Descending Sort similarity of Neighbor's book)
Output: List = List of recommend book
 1: procedure RecommendFiltering(I, x, Ix, L)
2. price = price of book
3: for each x in U
4:    if ( Rc = 1 ) then
5:        if Ix > median(Ic) then
              drop (L)
6:        endif
7:    endif
8: end for
9: for each I in I do
10:    if (price < max(price of x) or (price > max(price of x)
11:       drop(L)
12:    end if
13: end for
14: Return List(L)
15: end procedure
```

4. 실험 및 결과

4.1 실험 및 평가 방법

본 연구에서 제시한 추천방법의 성능을 평가하기 위해 2014년 1월부터 12월까지 수집된 국내 대형 온라인 서점의 고객 트랜잭션 데이터 534,883건을 사용하였다. 또한 도서의 정보를 추가적으로 얻기 위해 크롤링(crawling)을 진행하여 온라인 서점 사이트에서 제공하는 각 도서의 '책 소개'와 '목차' 정보를 수집하였다.

반복 실험을 통해 신뢰도가 높은 결과를 산출하기 위해 본 연구에서는 5-fold cross validation을 진행하였으며 5번의 정확도 실험 결과를 평균하여 추천 성능의 지표로 사용하였다. 또한 실험의 정확성을 높이기 위해 반복적으로 출간되는 잡지, 학습지를 제외하였으며, 판매나 구입횟수가 극히 적은 도서나 사용자는 추천 성능을 저하시키고 추천 결과를 왜곡함에 따라(임일 2015), 10회 이하로 판매되거나 구입횟수가 10회 이하인 사용자는 배제하였다. 그에 따라 정제된 데이터의 훈련용 집합(Train set)에는 9,639명의 회원과 36,760권의 도서가 포함되었다. 실험도구로는 한글 형태소 분석을 위해 꼬꼬마 형태소 분석기를 사용하였으며, Python 프로그래밍 언어를 사용하여 자료처리와 결과 분석을 하였다.

추천 시스템 성능을 평가하기 위해 대표적으로 사용되는 지표는 정확도(Precision)와 재현율(Recall)이다(Billsus and Pazzani 1999; Herlocker et al. 2004; Sarwar et al. 2001). 정확도는 추천한 전체 도서 개수 중 실제 구매한 도서의 개수를 나눈 비율이고, 재현율은 실제 구매한 전체 도서 개수 중에서 추천된 도서의 개수를 나눈 비율이다. 하지만 정확도와 재현율은 서로 반대의 관계에 있어 재현율을 높이기 위해서 추천하는 아이템의 수를 늘리면 정확도가 낮아지고 정확도를 높이기 위해 추천하는 아이템의 수를 줄이면 재현율이 낮아지는 문제가 발생된다(Sarwar et al. 2001). 따라서, 본 연구에서는 이러한 한계를 극복하기 위해 식(5)와 같이 정확도와 재현율을 동시에 고려하는 F-1지표(F-1 Measure)를 추천 시스템의 성능 평가 지표로 사용하였다(김경재·안현철 2009; Billsus and Pazzani 1999; Herlocker et al. 2004; Sarwar et al., 2001).

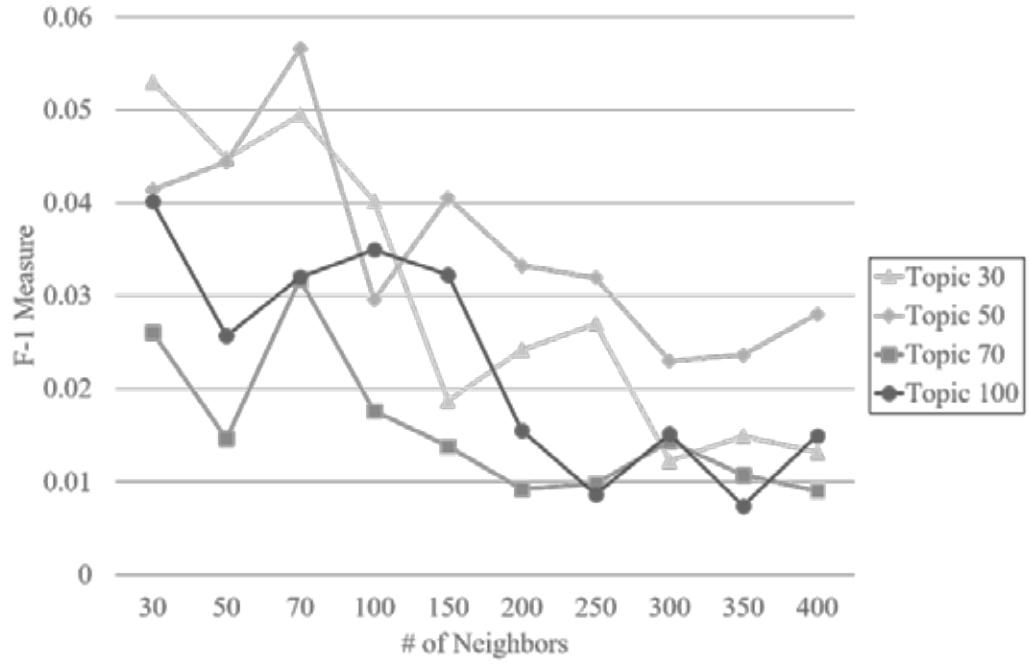
$$\begin{aligned} \text{정확도(Precision)} &= \frac{\text{추천 아이템} \cap \text{실제 구매한 아이템}}{\text{추천 아이템 수}} \\ \text{재현율(Recall)} &= \frac{\text{추천 아이템} \cap \text{실제 구매한 아이템}}{\text{실제 구매한 아이템 수}} \\ \text{F}-1\text{지표(F}-1\text{ Measure)} &= \frac{2\times Precision\times Recall}{Precision+Recall} \end{aligned} \quad (5)$$

4.2 실험 결과

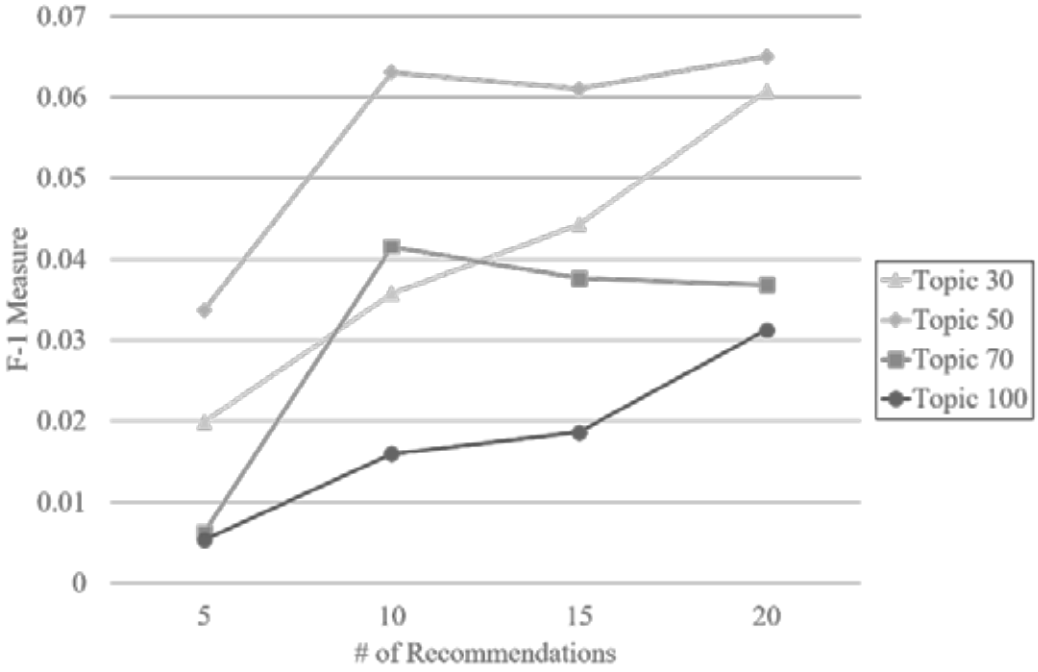
F-1지표를 사용하여, BCBCF에 대한 실험을 다각도로 진행하였다. 또한 BCBCF에서 토픽 모델링의 효과를 파악하기 위하여 도서의 최신성과 가격 수용도만을 적용한 방법도 비교하였다.

먼저, 추천 시스템의 성능은 토픽의 수와 이웃 집합의 크기, 추천 도서의 수에 따라 달라질 수 있다. 그에 따라 토픽의 수는 30, 50, 70, 100으로 변화시키며 추천 시스템의 성과를 측정하였으며 이웃 집합의 크기는 30~400명으로 변화시키며 그 성능을 측정하였다. 토픽 수와 이웃 집합의 크기 변화에 따른 실험 결과는 <그림 5>와 같다.

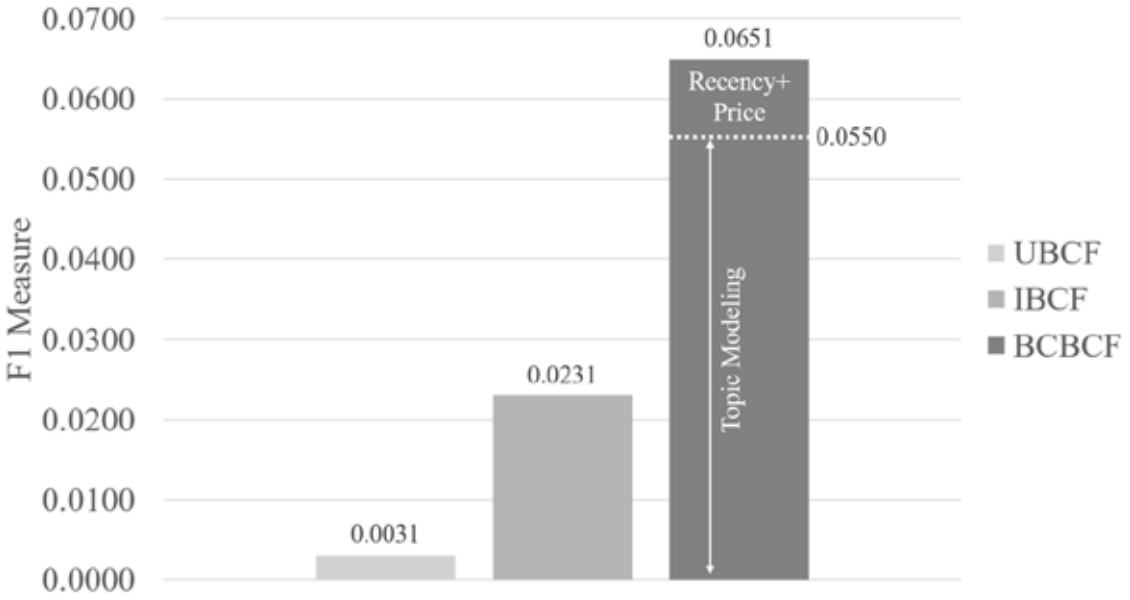
<그림 5>의 결과에서 토픽의 수를 30, 50, 70, 100개로 적용하고 각 이웃 집합의 크기를 변화하여 측정한 결과, 토픽을 50개로 하였을 때의 추천의 F-1지표가 다른 토픽 수보다 전체적으로 높게 나타났다. 또한 토픽 50의 결과 중 이웃 집합의 크기를 70으로 했을 때까지 성능이 향상되다가 이후로 점차 낮아지는 것을 알 수 있다. 이러한 결과는 토픽을 30, 70으로 했을 때나 토픽을 100으로 했을 때에도 마찬가지의 추세를 나타낸다. 위 결과에 따라 토픽은 50개로, 이웃 집합의 크기를 70으로 할 때 제안한 방법론의 성능이 가장 높다고 판단하였다.



<그림 5> 토픽별 네이버 수에 따른 추천 성과 비교



<그림 6> 토픽별 추천 도서수에 따른 추천 성과 비교



<그림 7> 추천 모델별 성과 비교

다음으로 추천 도서의 수를 결정하기 위해 이웃 집합의 크기를 70으로 고정시킨 후 추천 도서의 수를 5, 10, 15, 20개로 증가시키면서 성능을 비교하였다. 그 결과는 <그림 6>과 같다.

<그림 6>에 따르면 대부분 추천 도서수가 20일 때 가장 추천의 성능이 높아진다. 따라서 본 연구에서는 이러한 실험 결과를 바탕으로 추천 성능이 가장 높게 나오는 토픽 50, 이웃 집합 크기 70, 추천 도서 수 20으로 설정하여 나머지 실험을 진행하였다. <그림 7>은 실험을 통해 설정된 값을 기반으로 각 방법론을 비교한 결과이다.

<그림 7>에서 첫 번째와 두 번째 막대그래프는 사용자-도서를 중심으로 분석한 기존의 단순 협업 필터링(UBCF, IBCF) 방법이며 세 번째 그래프는 BCBCF에 대한 결과이다. 실험 결과 본 연구에서 제안한 BCBCF가 UBCF(0.0031), IBCF(0.0231)보다 더 높은 성과를 나타내었다. 특히 토픽 모델링만 적용하여도 0.055의 성과를 나타내어 기존의 협업 필터링 기법 중 높은 성과를 보인 IBCF에 비해 238%의 성능 향상이 이루어졌으며 최신성과 가격까지 포함한 경우 281% 향상되어 본 연구에서 제안한 방법이 기존의 협업 필터링 기법이 가지는 한계를 토픽 모델링과 최신성과 가격의 반영을 통해 일정 부분 해결했다고 할 수 있다.

5. 결론

사용자의 지식 및 정보 과부하를 줄여 서비스에 대한 만족도를 높이기 위해 다양한 분야에서 추천 시스템이 시도되고 있다. 대표적으로 활용되는 추천 시스템은 협업 필터링으로 지식의 공유 측면에서 다른 사용자의 정보를 기반으로 목표 사용자의 선호를 추론하여 높은 성과를 보이고 있다. 본 연구는 온라인 도서 추천에 특화된 방식으로, 저장되어 있는 사용자 구매 DB와 도서의 목차와 책 소개 내용에 LDA 토픽모델링 방식

을 적용한 결과를 결합하여 토픽 주제에 대한 사용자 선호와 도서구입에 대한 최신성과 가격 수용도 등의 사용자 선호를 반영하여 새로운 시스템을 제안하였다는 점에서 의의가 있다.

제안한 시스템은 기존의 협업 필터링 기법보다 높은 성과를 보여 협업 필터링이 가지는 한계를 최신성과 가격에 대한 반영, 토픽 모델링의 도입으로 충분히 해결 가능함을 보였으며 추가적으로 다음과 같은 효과를 관찰할 수 있었다. 먼저, 추천시스템을 개발하고자 할 때 토픽 모델링을 사용하여 희박성을 낮출 수 있었다. 일반적으로 희박성이 낮아지면 추천 성능은 좋아진다고 알려져 있다 (Huang et al. 2004). 사용자-도서 매트릭스에서의 희박성이 99.9580% 이고, 도서-토픽 매트릭스에서의 희박성이 95.8863%로 나타났는데, 두 매트릭스를 결합한 후 사용자-토픽 매트릭스의 희박성은 토픽을 50개로 적용하였을 때 63.0070%으로 낮게 나타났다. 그 결과 이웃 집합의 탐색이 용이해졌다. 두번째로 본 연구에서는 토픽과 도서의 최신성과 가격 수용도를 적용하여 추천리스트를 정교화하였다. 각 카테고리에 속하는 도서가 주문된 날짜와 출판된 날짜의 차이를 계산하였을 때, 25개 카테고리 중 19개의 카테고리 도서에서 3년 이내의 날짜 차이를 보였으며, 모든 카테고리의 기간 중위수가 평균값보다 짧은 기간차이를 나타내어, 대부분의 사용자가 도서의 최신성을 중시하고 있음을 확인할 수 있었다. 도서의 최신성과 사용자가 과거에 구입한 가격범위를 가격 수용도로 적용하여 사용자를 위해 보다 정교하게 추천리스트를 생성함에 따라 성과 향상에 많은 기여를 한 것으로 판단된다.

전통적인 협업 필터링 기법들과 비교 실험 결과에서는 본 연구에서 제안한 방법이 더 높은 성능을 보였으며 토픽 모델링만을 사용하는 것보다 최신성과 가격을 반영하였을 때 더 높은 성능을 보였다. 그에 따라 제안한 알고리즘이 도서 추천 시스템의 성능을 높이는 데 기여했다고 판단할 수 있다.

하지만 본 연구는 다음과 같은 한계를 지닌다. 먼저, 본 연구에서는 추천 리스트 생성 과정에서 최신성 선호와 가격 수용도에 적합하지 않은 도서를 단순히 제외하는 방법을 사용하였다. 토픽 모델링의 결과처럼 유사도 판단에서 최신성과 가격 수용도에 가중치를 두어 포함시킬 수 있다면 또 다른 효과를 가져올 수 있을 것이다. 그리고, 가격 수용도에 있어서도 이전에 구입한 개인별 도서 가격의 최저와 최고값만을 단순 적용하였다. 하지만 특정 카테고리만 구입한 사용자의 경우 가격 차이가 적을 수 있음에 따라 사용자의 카테고리별 가격 수용도를 적용하여 검증해볼 필요가 있다.

참고문헌

[국내 문헌]

1. 김경재, 안현철 2009. “개인화된 추천시스템을 위한 사용자-상품 매트릭스 축약기법,” *Journal of Information Technology Applications & Management* (16:1), pp. 97-113.
2. 김민정, 조윤호 2015. “빅데이터 기반 추천시스템 구현을 위한 다중 프로파일 앙상블 기법,” *지능정보연구* (21:4), pp. 93-110.
3. 박상현, 문현실, 김재경 2017. “토픽 모델링에 기반한 온라인 상품 평점 예측을 위한 온라인 사용 후기 분석,” 한국IT서비스학회지 (16:3), pp. 113-125.
4. 손지은, 김성범, 김현중, 조성준 2015. “추천 시스템 기법 연구동향 분석,” *대한산업공학회지* (41:2), pp. 185-208.
5. 양승준, 이보연, 김희웅 2016. “토픽모델링 기반 행복과 불행 이슈 분석 및 행복 증진 방안 연구,” *지식경영연구* (17:2), pp. 165-185.
6. 이세윤, 이선, 이정우 2006. “인터넷 콘텐츠 서비스 속성과 가격반응함수의 상관관계 분석,” *한국IT서비스학회* 2006년도 추계학술대회 논문집, pp. 80-85
7. 임일 2015. R을 이용한 추천 시스템, 서울:카오스북.
8. 전지은, 이충권 2014. “온라인 판매자들의 가격조정에 관한 연구,” *한국전자거래학회지* (19:3), pp. 143-158
9. 차윤정, 이지혜, 최지은, 김희웅 2015. “소셜미디어 토픽모델링을 통한 스마트폰 마케팅 전략 수립 지원,” *지식경영연구* (16:4), pp. 69-87
10. 최준연, 이석기, 조영빈 2013. “추천 시스템의 예측 정확도 향상을 위한 고객 평가정보의 신뢰도 활용법,” 한국콘텐츠학회논문지 (13:7), pp. 379-385.
11. 현윤진, 한희준, 최희석, 박준형, 이규하, 곽기영, 김남규 2013. “텍스트 분석을 활용한 국가 현안 대응 R&D 정보 패키징 방법론,” *Journal of Information Technology Applications & Management* (20:3), pp. 231-257.
12. 현윤진, 김남규, 조윤호 2015. “토픽 분석을 활용한 관심 기반 고객 세분화 방법론,” *정보기술연구* (22:1), pp. 77-93

[국외 문헌]

1. Adomavicius, G. and 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), pp. 734-749.
2. Bacos, Y. J. 1997. “The Emerging Role of Electronic Marketplace on the Internet,” *Communication of ACM* (41:8), pp. 35-42.
3. Bakos, J.Y. 1997. “Reducing buyer search costs: implications for electronic marketplaces’’, *Management Science* (43:12), pp. 1676-92.
4. Balabanović, M. and Shoham, Y. 1997. “Fab: content-based, collaborative recommendation,” *Communications of the ACM* (40:3), pp. 66-72.
5. Barragáns-Martínez, A. B., Costa-Montenegro, E., Burguillo, J. C., Rey-López, M., Mikic-Fonte, F. A. and Peleteiro, A. 2010. “A hybrid content-based and item-based collaborative filtering approach to recommend TV programs enhanced with singular value decomposition,” *Information Sciences* (180:22), pp. 4290-4311.
6. Billsus, D. and Pazzani, M. J. 1999. “A hybrid

user model for news story classification," in *UM99 User Modeling*, pp. 99-108.

7. Blei, D. M., & Lafferty, J. D. 2009. "Topic models," *Text mining: classification, clustering, and applications* (10:71), pp. 34.
8. Bobadilla, J., Ortega, F., Hernando, A. and Gutiérrez, A. 2013. "Recommender systems survey," *Knowledge-based systems* (46), pp. 109-132.
9. Brynjolfsson, E., Hu, Y. and Simester, D. 2011. "Goodbye pareto principle, hello long tail: The effect of search costs on the concentration of product sales," *Management Science* (57:8), pp. 1373~1386.
10. Burke, R. 2002. "Hybrid recommender systems: Survey and experiments," *User modeling and user-adapted interaction* (12:4), pp. 331-370.
11. Cho, Y. H. and Kim, J. K. 2004. "Application of Web usage mining and product taxonomy to collaborative recommendations in e-commerce," *Expert systems with Applications* (26:2), pp. 233-246.
12. Ding, Y., Li, X. and Orlowska, M. E. 2006. "Recency-based collaborative filtering," in *Proceedings of the 17th Australasian Database Conference* (49), pp. 99-107.
13. Griffiths, T. L. and Steyvers, M. 2004. "Finding scientific topics," in *Proceedings of the National academy of Sciences* (101:1), pp. 5228-5235.
14. Haucap, J. and Heimeshoff, U. 2014. "Google, Facebook, Amazon, eBay: Is the Internet driving competition or market monopolization?," *International Economics and Economic Policy* (11:1), pp. 49-61.
15. Herlocker, J. L., Konstan, J. A., Terveen, L. G. and Riedl, J. T. 2004. "Evaluating collaborative filtering recommender systems," *ACM Transactions on Information Systems* (TOIS) (22:1), pp. 5-53.
16. Hofmann, T. 1999. "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 50-57.
17. Huang Z, Chen H, Zeng D. 2004. "Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering," *ACM Transactions on Information Systems* (TOIS) (22:1), pp.116-142.
18. Li, L., Zheng, L., Yang, F. and Li, T. 2014. "Modeling and Broadening Temporal User Interest in Personalized News Recommendation," *Expert Systems with Applications* (14:7), pp. 3168-3177.
19. López, V., del Río, S., Benítez, J. M. and Herrera, F. 2015. "Cost-sensitive linguistic fuzzy rule based classification systems under the MapReduce framework for imbalanced big data," *Fuzzy Sets and Systems* (258), pp. 5-38.
20. Merve, A. and Arslan, A. 2009. "A collaborative filtering method based on artificial immune network", *Expert Systems with Applications* (36:4), pp. 8324~8332.
21. Mimno, D., Wallach, H. and McCallum, A. 2008. "Gibbs sampling for logistic normal

topic models with graph-based priors," In *NIPS Workshop on Analyzing Graphs,* pp.1-8.

22. Moon, H. S., Yoon, J. H., Choi, I. Y., Kim, J. K. 2017. "An Exploratory Study of Collaborative Filtering Techniques to Analyze the Effect of Information Amount," *Asia Pacific Journal of Information Systems* (27:2), pp. 126-138.
23. Park, C.H. and Kim, Y.G. 2003. "Identifying key factors affecting consumer purchase behavior in an online shopping context," *International Journal of Retail & Distribution Management* (31:1), pp. 16-29
24. Pazzani, M.J. and Billsus, D. 2007. "Content-based Recommendation Systems", in *The adaptive web*, P. Brusilovsky, A. Kobsa, and W. Nejdl (eds.), Berlin:Springer, pp. 325-341.
25. Peterson, R.A., Balasubramanian, S. and Bronnenberg, B.J. 1997. "Exploring the implications of the Internet for consumer marketing", *Journal of the Academy of Management Science* (25:4), pp. 329-346.
26. Ricci, F., Rokach, L. and Shapira B. 2011. *Introduction to recommender systems handbook*, MA: Springer US.
27. Salton, G. and Buckley, C. 1988. "Term-weighting approaches in automatic text retrieval," *Information processing & management* (24:5), pp. 513-523.
28. Sarwar, B., Karypis, G., Konstan, J. and Riedl, J. 2001. "Item-based collaborative filtering recommendation algorithms, " in *Proceedings of the 10th International Conference on World Wide Web*, pp. 285-295.
29. Sheydaei, N., Saraee, M. and Shahgholian, A. 2015. "A novel feature selection method for text classification using association rules and clustering," *Journal of Information Science* (41:1), pp. 3-15.
30. Song, M. and Kim, S. Y. 2013. "Detecting the knowledge structure of bioinformatics by mining full-text collections," *Scientometrics* (96:1), pp. 183-201.
31. Sugiyama, K., Hatano, K. and Yoshikawa, M. 2004. "Adaptive web search based on user profile constructed without any effort from users," in *Proceedings of the 13th international conference on World Wide Web*, pp. 675-684.
32. Wilson, J., Chaudhury, S. and Lall, B. 2014. "Improving collaborative filtering based recommenders using topic modelling," in *Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, pp. 340-346.
33. Wu, Y. H., and Chen, A. L. 2000. "Index structures of user profiles for efficient web page filtering services," in *Proceedings of the 20th International Conference on Distributed Computing Systems*, pp. 644-651.
34. Xin, J., Wang, Z., Qu, L. and Wang, G. 2015. "Elastic extreme learning machine for big data classification," *Neurocomputing* (149), pp. 464-471.
35. Zhang, Z., Lin, H., Liu, K., Wu, D., Zhang, G. and Lu, J. 2013. "A hybrid fuzzy-based

personalized recommender system for telecom products/services," *Information Sciences* (236), pp. 117-129.

36. Zhou, X. and Wu, S. 2016. "Rating LDA model for collaborative filtering," *Knowledge-Based Systems* (110), pp. 135-143.

저 자 소 개

정 영 진 (Jung, Youngjin)

현재 국민대학교 데이터사이언스학과 박사과정에 재학 중이며, 서울문화고등학교에 재직 중이다. 동국대학교 전자계산학과에서 학사 및 교육석사학위를 취득하였다. 주요 관심분야는 데이터마이닝, 추천시스템, 교육데이터분석 등이다.

조 윤 호 (Cho, Yoonho)

현재 국민대학교 경영학부 빅데이터경영통계전공 교수로 재직 중이다. 서울대학교 계산통계학과를 졸업하고, KAIST 경영정보공학과에서 석사, KAIST 경영공학과에서 박사학위를 취득하였으며, LG전자(주)에서 6년간 주임연구원으로 재직하였다. 주 연구분야는 머신러닝, 딥러닝, 비즈니스애널리틱스, 추천시스템, 소셜네트워크분석, 고객관계관리 등이다.