Information Systems Review
Vol. 23, No. 1 February 2021 http://dx.doi.org/10.14329/isr.2021.23.1.155

리뷰 데이터 마이닝을 이용한 하이브리드 추천시스템 개발: Amazon Kindle Store 데이터 분석사례

Development of Hybrid Recommender System Using Review Data Mining: Kindle Store Data Analysis Case

장 예 화 (Yihua Zhang) 경희대학교 일반대학원 빅데이터응용학과 석사과정
이 청 용 (Qinglong Li) 경희대학교 일반대학원 빅데이터응용학과 석사과정
최 일 영 (Ilyoung Choi) 경희대학교 경영대학원 & AI 경영연구센터
김 재 경 (Jaekyeong Kim) 경희대학교 경영대학 & 빅데이터응용학과, 교신저자

요 약

최근 온라인 상품 구매의 증가로 인해 사용자의 선호에 맞는 상품을 추천해주는 시스템이 지속적으로 연구되고 있다. 추천 시스템은 사용자들에게 개인화된 상품 추천 서비스를 제공하는 시스템으로 사용자가 상품에 남긴 평점을 이용한 협업 필터링(Collaborative Filtering)이 가장 널리 쓰이는 추천 방법이다. 협업 필터링에서 상품 간의 유사도 계산은 시간이 많이 소요되는데, 특히 리뷰 데이터와 같은 빅데이터를 사용할 경우 더욱 많은 시간을 소요한다. 그래서 본 연구에서는 리뷰 데이터 마이닝을 이용하여 상품 간의 유사도 계산을 빠르게 수행할 수 있으면서 정확도를 높일 있도록 2단계(2-Phase) 방법을 이용한 하이브리드 추천시스템 방식을 제안한다. 이를 위해 온라인 전자책 상거래 상점인 아마존 킨들 스토어(Amazon Kindle Store)의 약 98만 개의 온라인 소비자 평점과 리뷰 데이터를 수집하였다. 실험 결과 본 연구에서 제안한 사용자의 평점과 리뷰를 단계적으로 반영한 하이브리드 추천 방식이 전통적인 추천 방식과 비교하여 추천 시간은 비슷하였으나 높은 정확도를 나타내는 것을 확인하였다. 따라서 제안한 방법을 사용하면 사용자가 선호하는 상품을 빠르고 정확하게 추천함으로써 고객의 만족을 높여서 기업의 매출 증대에 기여할수 있을 것으로 기대된다.

키워드 : 리뷰 데이터 마이닝, 텍스트마이닝, 추천시스템, 협업 필터링

Ⅰ. 서 론

최근 정보통신기술의 발전과 스마트 기기의 대중화로 인해 다양한 종류의 데이터가 급격히 증가하고 있다. 하지만 기하급수적으로 증가하는 정보로 인해 사용자들은 정보 과부화 문제에 직면하여 필요한 상품 또는 서비스를 선택하는데 많은 시간이 필요하고 어려움에 직면하고 있다(Kim *et al.*, 2010; Lee *et al.*, 2020). 이와 따라 사용자에게 맞춤형 정보를 제공할 수 있는 개인화 추천 서비스에 대한 중요성이 대두되고 있다. 대표적으로 Amazon (Linden *et al.*, 2003), Netflix(Bennett and Lanning,

2007), YouTube(Covington *et al.*, 2016) 등 세계적인 전자상거래 기업들은 사용자의 구매기록을 기반으로 개인화 서비스를 제공하여 기업의 지속가능한 경쟁력을 꾸준히 강화하고 있다.

기존 추천 시스템 연구에서는 평점 등과 같이 사용자의 선호도를 정량적으로 나타내는 명시적 데이터(Explicit Data)와 구매 혹은 클릭 여부를 나타내는 암묵적 데이터(Implicit Data)를 사용하여 사용자에게 적합한 상품 또는 서비스를 제공했다(Cho and Kim, 2004; Su and Khoshgoftaar, 2009). 최근 연구에서는 이와 같은 정량적인 데이터만을 사용하여 추천 시스템을 구축하면 추천 성능이 떨어질 수 있다는 문제가 제기되고 있다(Zhang *et al.*, 2014). 예를 들어, 사용자 A와 사용자 B는 같은 상품을 구매하고 평점을 동일하게 4점을 부여했을 때 해당 상품에 대한 사용자의 정량적인 선호도는 같다고 볼 수 있다. 하지만 사용자 A는 배송 및 사은품 서비스 등 만족하고, 사용자 B는 배송 서비스에 만족하지 못하지만 상품의 품질과 디자인에 만족했다. 사용자 A와 사용자 B는 동일한 상품에 대한 정량적인 평점은 같지만, 해당 상품에 대한 정성적인 선호도는 같다고 보기는 어렵다. 따라서 사용자가 상품을 구매하거나 선호하는 이유 등을 파악할 수 있는 정성적인 선호도를 고려하지 않고, 단순히 정량적인 평점이나 구매 여부 등 데이터만을 고려하여 상품을 추천하는 것은 추천의 정확도를 떨어뜨리는 요소로 될 수 있다(전병국, 안현철, 2015; 현지연 등, 2019).

최근에는 이러한 기존 연구의 한계를 극복하기 위해 개인화 추천 서비스를 제공할 때 다양한 유형의 데이터들이 사용되고 있다. 대표적으로 많이 사용되는 것이 구매동기, 구매후기 등 정성적인 선호도를 나타낼 수 있는 리뷰 데이터이다(Zheng *et al.*, 2017; Batmaz *et al.*, 2019). 리뷰 데이터는 사용자가 특정 상품에 대하여 구체적이고 신뢰할 수 있는 선호도 정보를 포함하고 있기에 개인화 추천 서비스 제공할 때 유용하게 사용할 수 있다(전병국, 안현철, 2015; 현지연 등, 2019). 전병국, 안현철(2015)은 개인화 추천 서비스에서 사용자 간의 유사도를 계산할 때 리뷰 데이터의 유사도를 추가로 고려하여 보다 정교하게 사용자 간의 유사도를 산출하는 추천 방법론을 제안하였다 현지연 등(2019)은 개인화 맞춤형 서비스를 제공할 때 사용자 리뷰를 정량적인 선호도 정보로 변환하여 추천 시스템에 직접 반영하는 새로운 추천 방법론을 제안했다. 이와 같은 추천 방법론들은 개인화 추천 서비스를 제공할 때 모든 평점과 리뷰 데이터를 사용하여 유사도를 계산하기 때문에 모델 확장성(Scalability) 문제가 존재하고 있다(Herlocker *et al.*, 2004; Jannach *et al.*, 2010). 다시 말해 추천 시스템을 구축할 때 모든 데이터를 사용하여 연산을 수행하기에 많은 시간과 비용이 소모되는 문제점이 발생한다. 개인화 추천 서비스의 기본적인 아이디어는 실시간으로 변화되는 정보를 통해 신속하고 정확하게 사용자가 선호할 만한 상품을 추천해준다. 즉, 즉각적인 피드백 구조가 유지되기 위해서 알고리즘의 빠른 계산 속도는 매우 중요한 사항이다(Bobadilla *et al.*, 2013; Ricci *et al.*, 2011).

본 연구에서는 이와 같은 기존 연구의 한계점을 개선하기 위해 평점과 리뷰를 순차적으로 고려하는 추천 방법론을 제안하고자 한다. 본 연구에서 제안하는 방법론은 다음과 같다. 첫째, 먼저 사용자 평점만을 사용하여 유사도를 계산하여 상위 N개 추천 목록을 생성한다. 둘째, 상위 N개 추천 목록의 상품 리뷰 데이터에 TF-IDF 기법을 적용하여 유사도를 계산하고 이를 추천 시스템에 반영하여 최종 상위 N개 추천 목록을 생성한다. 마지막으로, 본 연구에서는 평점과 리뷰를 순차적으로 고려하는 추천 방법론의 추천 성능을 평가하기 위해 추천 시스템 연구에서 널리 사용되고 있는 F1-Score 측정지표를 사용하여 기존 연구와 추천 성능을 비교했다. 실험 결과 본 연구에서 제안한 추천 방법론이 기존의 추천 방법론과 비교했을 때 추천 정확도와 추천 소요시간 모두 일정하게 개선을 보여주고 있다.

본 연구의 구성은 다음과 같다. 제Ⅱ장에서는

협업 필터링과 텍스트 마이닝에 대한 이론적 배경을 간략하게 서술하고, 관련 연구에 대하여 살펴본다. 제Ⅲ장에서는 본 연구에서 제안하는 추천 방법론에 대해 자세하게 설명하고, 데이터 수집, 실험설계와 성능평가 방법을 기술한다. 제Ⅳ장에서는 제안한 추천 방법론의 실험 결과를 기술한다. 마지막으로 제Ⅴ장에서는 본 연구의 결론을 기술하고, 본 연구의 한계점 및 향후 연구계획에 대하여 기술한다.

Ⅱ 이론적 배경

2.1 협업 필터링

협업 필터링(Collaborative filtering, CF) 기법은 전자상거래 분야에서 개인화 추천 서비스를 제공할 때 가장 널리 사용되고 있으며, 현재까지 가장 성공적인 추천 기법 중 하나로 알려져 있다(Goldberg *et al.*, 1992; Kim *et al.*, 2005; Kim *et al.*, 2018). CF 기법의 기본적인 아이디어는 사용자의 선호도를 예측하기 위해 유사도 측정을 통해 비슷한 유사한 이웃 사용자를 선정하는 것이다(Choi *et al.*, 2019a; Choi *et al.*, 2019b; Schafer *et al.*, 2007). CF 기법은 사용자 - 사용자 간의 유사도를 계산하여 선호도를 예측하는 사용자 기반 CF 기법, 상품-상품 간의 유사도를 계산하여 선호도를 예측하는 상품 기반 CF 기법으로 구분된다(Bobadilla *et al.*, 2013; Ricci *et al.*, 2011). 대표적으로 Amazon(Linden *et al.*, 2003), Netflix(Bennett and Lanning, 2007) 등 전사상거래 기업은 개인화 추천 서비스를 제공할 때 주로 상품 기반 CF 기법을 사용하고 있다. 그 이유는 사용자 기반 CF 기법은 사용자가 상품을 평가할 때 다른 상품에 대하여 추천을 제공해야 하나 평가할 때 마다 유사도 정보를 반영하는 것은 어렵다는 한계점이 존재한다. 따라서 대부분의 경우 사용자에 비해 상품수가 적기 때문에 상품 간의 상호관계를 발견할 수 있는 확률이 높기에 추천 제공할 때 더 적합하다.

CF 기법은 개인화 추천 서비스 분야에서 큰 성공에도 불구하고 다음과 같은 한계점을 가지고 있다. 첫 번째는 데이터 희소성 문제점(Sparsity Problem)이 존재한다(Herlocker *et al.*, 2000; Shani and Gunawardana, 2011). 사용자의 정확한 선호도를 예측하기 위해서는 상품에 대한 선호도를 나타내는 평점 정보가 충분히 확보되어야 보다 정확한 추천 서비스 제공이 가능하다. 하지만 상품에 대한 평점 정보가 상품의 인기에 따라 편향되거나 구매가 이루어지지 않아 평점 정보가 없는 상품이 존재할 경우 사용자의 정확한 선호도를 예측하기 어렵다(Bobadilla *et al.*, 2013; Kim *et al.*, 2011b). 또한, 신규 사용자의 경우 선호도 정보가 아직 반영되지 않는 추천이 불가능한 콜드 스타트 문제점(Cold Start Problem)이 존재한다. 이러한 문제점을 해결하기 위해 이상기 등(2010)은 선호도 정보가 부족한 새로운 논문을 사용자에게 추천하기 위해 논문에 대한 평점 정보 대신 논문의 키워드와 사용자의 과거에 검색한 논문의 키워드를 비교하여 추천하는 방법을 제안하였다. 김병만 등(2004)은 새로운 고객의 프로필 정보를 사용하여 기존 사용자 간의 유사도를 계산하여 구매이력이 없는 새로운 고객의 선호도를 예측하여 추천하는 방법을 제안하였다. 두 번째는 확장성 문제점(Scalability Problem)이다. 사용자에게 개인화 추천 서비스를 제공할 때 추천 소요시간은 매우 중요한 사항이다. 최근에 사용 가능한 데이터 양이 급증하면서 추천 시스템의 연산에 많은 시간이 소모되면서 추천 소요시간이 감소되는 문제점이 나타나고 있다(Herlocker *et al.*, 2004; Jannach *et al.*, 2010). 이러한 문제점을 해결하기 위해 Park *et al.*(2015)은 유사도가 큰 순서대로 N명(개)의 사용자 혹은 상품을 선정하여 선호도를 예측하는 방법을 제안하였고 기존의 CF 기법보다 빠른 연산 속도를 나타내는 것을 확인하였다. 세 번째는 특이 취향 사용자 문제(Grey Sheep Problem)이다. CF 기법은 사용자가 상품의 유행에 따르거나 일정한 패턴을 가지고 있다는 가정하에 추천 제공하지만 일관성이 부족한

정보는 사용자의 선호도를 예측할 때 오히려 추천 성능을 저하시키는 문제가 발생한다(Herlocker *et al.*, 2000; Kim *et al.*, 2011a Shani and Gunawardana, 2011). 이와 같은 문제점을 해결하기 위해 Claypool *et al.* (1999)은 가중 평균값을 예측 값으로 사용하여 수정된 가중치를 반복적으로 적용하는 방법을 제안하였다. 네 번째는 개인화 추천 서비스를 제공할 때 평점, 리뷰, 상품 구매시간 등 다양한 정보가 필요하지만 CF 기법은 평점과 같은 기본적인 정량적인 정보만을 사용하는 한계점이 존재한다. 최근의 여러 연구에서 추천 제공할 때 사용자의 평점과 리뷰 정보를 동시에 고려할 때 추천 시스템의 성능이 향상되는 것을 확인하였다(전병국, 안현철, 2015; 현지연 등, 2019). 따라서 본 연구에서는 사용자의 선호도를 정교하게 예측하기 위해 평점과 리뷰를 순차적으로 고려하는 추천 방법론을 제안하고자 한다.

2.2 텍스트 마이닝

텍스트 마이닝(Text Mining)은 자연어로 구성된 대량의 비정형 텍스트 데이터에서 숨겨진 패턴 또는 관계를 추출하여 의미가 있고 활용가치가 높은 정보 또는 지식을 추출하는 분석기법이다(Berry and Castellanos, 2004). 그 중에서도 대표적으로 많이 사용되고 있는 비정형 데이터는 전자상거래에서 수집되고 있는 온라인 리뷰이다. 온라인 리뷰 데이터는 양의 방대하고 텍스트 형태로 구성되어 있기 때문에 전통적인 데이터 마이닝 기법으로 분석하기에는 한계가 존재한다(Gupta and Lehal, 2009).

개인화 추천 서비스 연구에서 온라인 리뷰를 분석할 때 대표적으로 많이 사용되는 기법은 TF - IDF(Term Frequency-Inverse document Frequency) 기법이다(Gupta and Lehal, 2009). TF-IDF 기법은 여러 문서로 이루어진 문서군에서 특정 단어가 문서 내에서 얼마나 중요한 것인가를 나타내는 방법으로, 문서의 핵심어를 추출하거나, 검색 엔진에서 검색 결과의 순위를 결정할 때, 혹은 문서 간 유사도를 산출하는 등 다양한 용도로 사용된다(Berry and Castellanos, 2004; Gupta and Lehal, 2009). 이성직, 김한준(2009)은 TF-IDF 기법을 적용하여 뉴스에서 추출한 단어 간의 교차비교분석을 통해 의미 없는 단어를 제거하여 단어 추출 성능을 개선하였다. 박대서, 김화종(2018)은 TF-IDF 기법에 단어 유사도를 적용한 의미 벡터를 결합하여 뉴스기사를 벡터화를 통해 뉴스 검색과 연관 뉴스 추천 연구를 수행하였다. 유은순 등(2015)은 사용자에게 맞춤형 도서 추천 서비스를 제공하기 위해 TF-IDF 기법을 적용하여 도서 본문 텍스트의 의미적 정보를 추천 시스템에 반영하는 방법론을 제안했다. 연구결과 추출 정확도가 기존 추천 방법론에 비해 좋은 성능을 보여주었다. 연다인 등(2020)은 스마트 스피커에 대해 텍스트 마이닝 기법을 적용하여 사용자가 실제 작성한 리뷰 데이터를 수집하여 스피커 사용자 경험 차원을 기반으로 분석 결과를 해석했다. 이를 통해 스마트 스피커 제조사에게 실무적으로 사용자 경험 강화를 위한 전략을 제안했다. Li and Zhang(2015)은 정보과부하 문제를 해결하기 위해 TF-IDF 기법 적용하여 사용자의 요구사항과 동기를 분석하는 알고리즘을 제안하였다. 본 연구에서는 키워드 추출 등 텍스트 분석 분야에서 널리 사용되고 있는 TF-IDF 기법을 적용하여 리뷰 데이터를 분석하여 사용자의 상품 선호도 예측하고자 한다.

Ⅲ. 연구설계 및 분석 방법

3.1 연구설계

기존의 추천 시스템 연구는 주로 사용자의 정량적인 평점 데이터를 사용하여 개인화 추천 서비스를 제공했다. 하지만 사용자의 구매동기, 구매후기 등 정성적인 선호도를 고려하지 않고 정량적인 평점만을 사용하여 추천 시스템을 구축하면 추천 성능이 떨어지는 문제가 발생한다. 본 연구에서는 이러한 기존 연구의 한계점을 개선하기 위해 평점

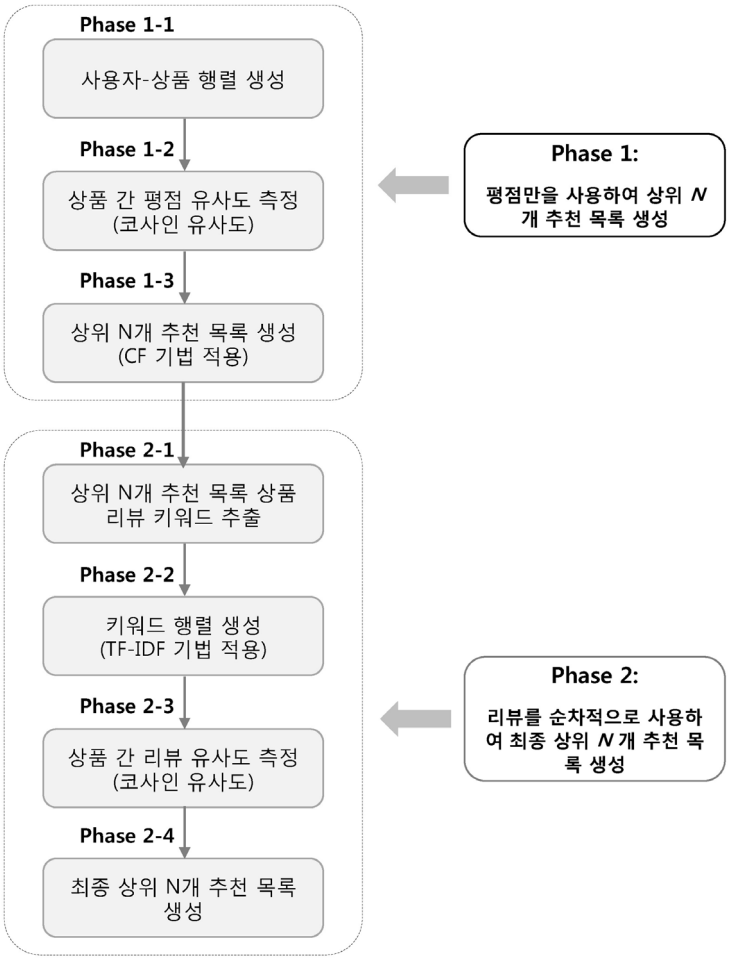
과 리뷰를 순차적으로 고려하는 추천 방법론을 제안하고자 한다. 본 연구에서 제안하는 추천 방법론은 <그림 1>과 같이 평점 기반 상위 N개 추천 목록 생성 단계와 리뷰 기반 상위 N개 최종 추천 목록 생성 단계로 구성된다.

Phase 1: 평점 기반 상위 N개 추천 목록 생성

본 연구에서 제안하는 추천 방법론의 첫 번째 단계에서는 사용자의 평점만을 사용하여 상위 N개 추천 목록을 생성한다. 첫 번째 단계는 구체적으로 사용자-상품 행렬 도출, 코사인 유사도를 사용한 유사도 측정과 상위 N개 추천 목록 생성 단계로 구성된다.

Phase 1-1: 사용자-상품 행렬 도출

해당 단계는 추천 대상 사용자의 선호도를 예측하기 위해 사용자 평점을 사용하여 <그림 2>와 사용자-상품 행렬 $m \times n$ 생성하며 사용자의 상품에 대한 선호도는 평점 $R=(r_{ij})$로 나타낸다. 여기서 r_{ij}는 $i(1 \le i \le m)$ 번째 사용자가 $j(1 \le j \le n)$ 번째 상품에 대한 평점을 나타내고 있으며 평점은 1부터 5사이의 값을 가진다.



〈그림 1〉 평점과 리뷰를 순차적으로 고려하는 추천 방법론

	아이템$_1$	아이템$_2$	…	아이템$_i$	…	아이템$_n$
사용자$_1$	4			5		4
사용자$_2$	5	3		3		
⋮						
사용자$_j$	3	1		3		2
⋮						
사용자$_m$	4	4		1		6

〈그림 2〉 사용자-상품 행렬 예시

Phase 1-2: 유사도 측정

유사도 측정 단계는 추천 시스템 구축 단계에서 중요한 단계이다. 기본적으로 사용자가 여러 상품에 같은 평점을 준다면 해당 상품들은 유사도가 높다고 볼 수 있다. Phase 1-1에서 도출한 사용자-상품 행렬의 평점 정보를 사용하여 식 (1)과 같이 코사인 기반 평점 유사도를 계산한다.

$$Sim(j,j') = \frac{(r_j \cdot r_{j'})}{(\|r_j\| \cdot \|r_{j'}\|)} \quad (1)$$

여기서 $Sim(j,j')$는 상품 j와 상품 j'간의 평점 유사도를 나타내고, r_j와 $r_{j'}$는 상품 j에 대해 사용자들이 평가한 점수의 집합 및 상품 j'에 대해 사용자들이 평가한 점수의 집합이다.

Phase 1-3: 상위 N개 추천 목록 생성

마지막 단계에서는 Phase 1-2에서 측정한 평점 유사도를 사용하여 구매 가능성 점수(Purchase Likelihood Score, PLS)를 계산하고 상위 N개 추천 목록을 생성한다. j번째 상품에 대한 사용자 i의 $PLS(i,j)$는 식(2)와 같이 계산한다.

$$PLS(i,j) = \frac{\sum_{n \in neighbors}(r(i,j) - \overline{r_j}) \times Sim(j,j_n)}{\sum_{n \in neighbors} Sim(j,j_n)} \quad (2)$$

여기서 $r(i,j)$는 사용자 i가 상품 j에 남긴 평점을 나타내고, $\overline{r_j}$는 상품 j에 대한 평점의 평균을 나타내며, j_n은 상품 j와 평점 유사도가 높은 N개의 상품을 의미한다. 또한 $Sim(j, j_n)$는 상품 j와 상품 j_n의 평점 유사도를 나타낸다. 따라서 PLS 값이 높다는 것은 이웃 상품이 구매가 될 확률이 높다는 것을 의미한다(박종학 등, 2009).

Phase 2: 리뷰 기반 상위 N개 추천 목록 생성

평점 기반 상위 N개 추천 목록 생성하고 리뷰 마이닝을 통해 추가적으로 리뷰 유사도 정보를 반영하여 최종 상위 N개 추천 목록을 생성한다. 두 번째 단계는 리뷰 전처리, 키워드 추출 및 행렬 생성, 리뷰 유사도 측정을 통해 최종 유사한 이웃 선정과 최종 상위 N개 추천 목록 생성 단계로 구성된다.

Phase 2-1: 리뷰 전처리

리뷰 데이터는 많은 단어를 포함하지만 모든 단어가 의미를 나타내는 것은 아니다. 예를 들어, ‘good’, ‘better’, ‘beautiful’ 등은 중요한 의미를 나타내는 반면 ‘He’, ‘my’, ‘are’ 등은 크게 중요한 의미를 나타내지 않는다. 리뷰 데이터에서 사용자의 선호도를 나타내는 중요한 단어만 추출하여 사용하면 데이터의 크기를 줄이고 모델의 효율성과 성능 개선에도 도움을 줄 수 있다. 따라서 본 연구에서는 <그림 3>과 같은 과정을 거쳐 전처리를 수행했다. 온라인에서 수집한 리뷰 데이터에는 HTML 태그나 다양한 부호 혹은 기호를 포함하고 있다. 이와 같은 노이즈 데이터(Noise Data)는 아

〈그림 3〉 리뷰 데이터 전처리 과정

무런 의미를 나타내지 않기 때문에 정규 표현식(Regular Expression)을 통해 제거했다. 조사, 접미사 등과 같은 불용어(Stopword)는 리뷰 데이터에 나타나는 빈도수는 많으나 실제 분석을 수행할 때 의미를 나타내지 않기에 제거했다. 또한, 리뷰 데이터에서 영어 대문자와 소문자를 통합하거나 비슷한 의미의 단어를 하나의 단어로 정규화를 하면 단어의 개수를 줄일 수 있는 또 다른 전처리를 수행하는 방법이다. 따라서 리뷰 데이터에서 포함된 비슷한 의미를 나타내는 단어의 개수를 줄이기 위해 어근 추출(Stemming) 작업을 진행하였다.

Phase 2-2: 키워드 추출 및 행렬 생성

Phase 2-1에서 전처리 과정을 거치고 이 단계에서는 식 (3)과 같이 TF-IDF 기법을 적용하여 키워드를 추출하고 각 키워드의 중요도에 가중치를 부여하여 키워드 행렬을 생성한다(Berry and Castellanos, 2004; Gupta and Lehal, 2009).

$$V(j,k) = TF(j,k) \times IDF(j,k) \tag{3}$$

여기서 $TF(j,k)$는 상품 j의 리뷰에서 키워드 k가 나타난 횟수를 의미하고, $IDF(j,k)$는 상품 j의 리뷰 집합에서 총 리뷰 문서의 수를 키워드 k를 포함한 리뷰 문서의 수로 나눈 뒤 로그를 취하여 얻은 값이다.

Phase 2-3: 리뷰 유사도 측정

이 단계에서는 Phase 2-2에서 생성한 키워드 행렬 정보를 사용하여 상품 간의 리뷰 유사도를 측정한다. 상품 간의 리뷰 유사도는 식 (4)와 같이 코사인 유사도를 사용하여 계산한다.

$$Sim(j,j') = \frac{(V_j \cdot V_{j'})}{(\|V_j\| \cdot \|V_{j'}\|)} \tag{4}$$

여기서 $Sim(j,j')$는 상품 j와 상품 j'의 리뷰 유사도를 나타내고, V_j와 $V_{j'}$는 각각 상품 j에 대한 리뷰 키워드의 중요도 집합을 의미한다.

Phase 2-4: 최종 상위 N개 추천 목록 생성

마지막 단계에서는 상품과 상품 간의 리뷰 유사도를 사용하여 PLS 점수를 계산하고 최종 상위 N개 추천 목록을 생성한다. 리뷰 유사도를 이용한 j번째 상품에 대한 사용자 i의 $PLS(i,j)$는 식 (5)와 같이 계산된다.

$$PLS(i,j) = \frac{\sum_{n \in neighbors} (V(j,k) - \overline{V_j}) \times Sim(j,j_n)}{\sum_{n \in neighbors} Sim(j,j_n)} \tag{5}$$

여기서 $V(j,k)$는 상품 j의 리뷰에 대한 키워드 k의 중요도를 나타내고, $\overline{V_j}$는 상품 j에 남겨진 리뷰 키워드들의 중요도에 대한 평균값을 나타내며, j_n은 상품 j와 리뷰 유사도가 높은 상품을 의미한다. 또한 j_n은 상품 j와 평점 유사도가 높은 N개의 추천 후보 상품을 나타내며, $Sim(j,j_n)$는 상품 j와 상품 j_n의 리뷰 유사도를 나타낸다.

3.2 분석방법

사용자가 구매한 상품에 대한 평점, 리뷰 데이터가 <표 1>과 같이 주어졌다고 가정한다. 본 연구에서 제안한 추천 방법론을 적용하여 “B00B34Q1O6” 상품을 구매한 사용자에게 추천 목록을 생성하는 과정을 아래와 같이 단계별로 구체적으로 설명할 수 있다.

〈표 1〉 평점 및 리뷰 데이터 예시

Item	User	Rating	Review
B00B34Q1O6	User 1	4	I really liked this story and has read it over a couple of times I di···
	User 2	3	This book had good potential, but sadly this didn‘t materialize. It···
	User 4	5	This was a great book from Leila and Tucker, Sarah, and let not f···
	User 5	3	I really enjoyed this story! It was so captivating I read it in one d···
	User 6	2	Huh the end of the book is weird what the hell happened the story··
	User 8	4	This book is just awesome I could not put it down. I not onlylove···
B00CC2HCK0	User 1	3	I thoroughly enjoyed this book and there was definitely enough···
	User 2	5	Love the concept, loved the two heros of the story, loved that she···
	User 3	4	This book is really good. I haven’t laughed out loud while read···
	User 5	4	I love that there are entire families and characters throughout the···
	User 6	4	OMG!!1 i can’t get enough of this book. i love it. i love how ···
	User 7	3	This is a good read. However, i was not able to finish the book b···
	User 8	5	Hot and sexy book. To bad men don’t turn into dragons. The st···
B00CC3QZ8E	User 2	1	This book had promise but fell short. Grammatical errors, numer···
	User 3	3	Flat characters with no development. The F-bomb dropped like ···
	User 4	4	Jessica Knapp loves her job as secretary to the CEO of the tradi···
	User 7	4	This short story was straight to the point. It was so fast paced tha···
B00CC5NG6Q	User 1	5	Men Don’t Cry by Boo Jackson was a damn good read, when I fi···
	User 2	3	This book could have very well been a full-length novel. Surpri···
	User 3	2	I wish it went into more detail but since I have been there, I kno···
	User 5	5	Hey Boo, this was truly an awesome read...sorry so late on the re···
	User 6	4	Having been involved as a volunteer in hospice, as well as being···
	User 8	3	Hospice is a difficult thing. It’s difficult for the person who is ail···
B00GMWTCOU	User 1	1	I was somewhat disappointed in this book. Think that maybe it ···
	User 2	5	I read the book as research not because I was going through the l···
	User 4	2	Written in 1953, this is noir, without a doubt. Loving that genre···
	User 6	2	If you are a lover of language, this book is a masterpiece. The aut···
	User 7	1	I have not received one of my fourteen free copy of the NYT so···
B00GN4LTWK	User 1	2	This book was so-so, even tho free. I don’t believe I will be sear···
	User 3	5	The writing was good and the plot moved quickly. I felt the relati···
	User 4	3	I didn’t get very far in this book so it’s a little unfair to say it stink···
	User 6	5	Cute story within the ″I need a hero″ niche. We all loved the knig···
	User 7	1	I didn’t like it, the first story was too graphic.....ugh. If I want to ···
	User 8	5	Jemma Leigh has been tormented by a stalker for months now. S···
B008C9HWRU	User 1	1	This may not be a fair review of the book, as I did not finish. I re···
	User 2	3	Right from the start the book pulls you in. So many secrets. You···
	User 4	5	Nice short read of interesting stories. I even went ahead and ord···
	User 5	3	It was an OK read although at sixty seven I found it a bit juvenile···
	User 6	4	Stephen Leather writes a great story. I read one of his full length···
	User 8	4	I liked this story overall. I thought the storyline was good. I was···
B00AS9KEGI	User 1	5	This format was very different from what I usually choose. The ···
	User 2	4	If you are a fan of Stephen Leather, you will enjoy these fun shor···
	User 3	5	Stephen Leather has composed quite a few books and short stori···
	User 6	5	This was a most enjoyable read! As good as Robert Parker’s Spe···
	User 7	2	I did not care for this book. It just did not make sense to me. Will···
	User 8	2	They are excerpts from other stories by the author and this book···

〈표 2〉 사용자 - 상품 평점 행렬 예시

	B00B34Q1O6	B00CC2HCK0	B00CC3QZ8E	B00CC5NG6Q	B00GMWTCOU	B00GN4LTWK	B008C9HWRU	B00AS9KEGI
User1	4	3		5	1	2	1	5
User2	3	5	1	3	5		3	4
User3		4	3	2		5		5
User4	5		4		2	3	5	
User5	3	4		5			3	
User6	2	4		4	2	5	4	5
User7		3	4		1	1		2
User8	4	5		3		5	4	2

먼저 상품 간의 유사도를 측정하기 위해 <표 1>의 데이터를 사용하여 사용자의 선호도 정보를 나타내는 평점 데이터를 추출하여 <표 2>와 같이 사용자-상품 평점 행렬을 생성한다.

생성된 사용자 - 상품 평점 행렬을 사용하여 상품 간의 유사도를 측정한다. <표 3>은 <표 2>의 평점 정보를 사용하여 상품 간의 유사도를 측정한 결과이다. 평점 유사도를 측정한 결과 상품 <B00B34Q1O6>와 유사도가 가장 높은 상품은 <B008C9HWRU>이라는 것을 <표 3>의 유사도 측정 결과를 통해 확인할 수 있다.

다음으로 식 (2)와 같이 PLS 점수를 계산하여 평점 기반 상위 N개 추천 목록을 생성하며, 결과는 <표 4>와 나타낸다. 7개의 상품 중에서 PLS 점수를 계산했을 때 상위 5개 상품 추천 목록은 <B008C9HWRU>, <B00CC5NG6Q>, <B00CC2HCK0>, <B00GN4LTWK>, <B00AS9KEGI>이다.

앞의 단계에서 평점 기반 때 상위 5개 상품 추천 목록을 생성하였다면, 다음 단계에서는 사용자의 리뷰 데이터를 사용하여 리뷰 기반 최종 상위 N개 상품 추천 목록을 생성한다. 먼저 전처리 과정을 거치고 다음에 TF-IDF 기법을 적용하여 먼저 키워드 행렬을 생성한다. <표 5>는 평점 기반 상위 5개 상품 추천 목록의 리뷰 데이터를 추출하여 리뷰 키워드에 각각 부여된 중요도를 나타낸 키워드 행렬을 나타내고 있다.

〈표 3〉 평점 유사도 결과 예시

	B00B34Q1O6	B00CC2HCK0	B00CC3QZ8E	B00CC5NG6Q	B00GMWTCOU	B00GN4LTWK	B008C9HWRU	B00AS9KEGI
B00B34Q1O6								
B00CC2HCK0	0.700							
B00CC3QZ8E	0.399	0.415						
B00CC5NG6Q	0.768	0.881	0.148					
B00GMWTCOU	0.628	0.612	0.443	0.505				
B00GN4LTWK	0.632	0.728	0.507	0.621	0.340			
B008C9HWRU	0.916	0.703	0.407	0.697	0.659	0.693		
B00AS9KEGI	0.565	0.849	0.419	0.782	0.629	0.767	0.519	

〈표 4〉 평점 기반 PLS 점수 결과 예시

Product	B008C9HWRU	B00CC5NG6Q	B00CC2HCK0	B00GN4LTWK	B00GMWTCOU	B00AS9KEGI	B00CC3QZ8E
PLS	0.323	0.624	0.842	0. 568	0.186	0.764	0.133

〈표 5〉 상품 - 리뷰 키워드 매트릭스

	Good	Fun	Hero	Issue	Short	Cheap	Bad	Gift	Necessary	Different
B00B34Q1O6		3.786	5.235			2.248		4.147		2.256
B008C9HWRU	3.786	5.786		4.147					1.693	1.693
B00CC5NG6Q				5.235	2.256			4.147		3.786
B00CC2HCK0	6.378				3.786			4.387	2.478	
B00GN4LTWK			3.786				1.396		2.367	
B00AS9KEGI		5.786	2.478			4.578		1.396		

〈표 6〉 상품 - 상품 리뷰 유사도

	B00B34Q1O6	B008C9HWRU	B00CC5NG6Q	B00CC2HCK0	B00GN4LTWK	B00AS9KEGI
B00B34Q1O6						
B008C9HWRU	0.368					
B00CC5NG6Q	0.387	0.418				
B00CC2HCK0	0.244	0.376	0.373			
B00GN4LTWK	0.510	0.102	0	0.140		
B00AS9KEGI	0.775	0.503	0.091	0.086	0.254	

〈표 7〉 리뷰 기반 PLS

Product	B00CC5NG6Q	B008C9HWRU	B00GN4LTWK	B00AS9KEGI	B00CC2HCK0
PLS	0.742	0.823	0.268	0.549	0.209

생성된 상품-키워드 행렬을 기반으로 상품 간의 유사도를 측정하면 <표 6>과 같은 결과를 나타낼 수 있다. 예를 들어 상품 <B00B34Q1O6>와 리뷰 유사도가 가장 높은 상품은 <B00AS9KEGI>라는 것을 알 수 있다.

리뷰 유사도를 사용한 최종 상위 N개 추천 목록은 생성하기 위해 PLS 점수를 계산하면 <표 7>과 같다. 따라서 PLS 점수를 사용하여 산출한 리뷰 기반 상위 3개 상품 추천 목록은 <B00CC5NG6Q>, <B008C9HWRU>, <B00AS9KEGI>이다.

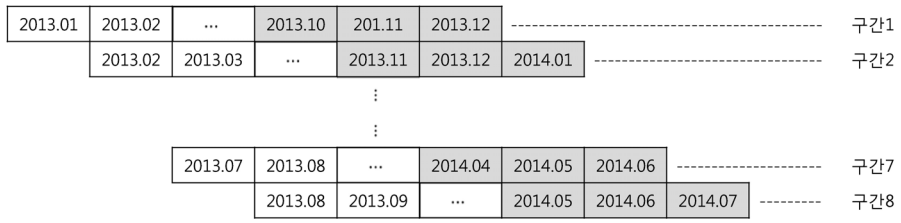
3.3 데이터 수집

본 연구에서 제안한 추천 방법론의 추천 성능을 검증하기 위해 전자상거래 사이트 Amazon에서 1996년 5월부터 2014년 7월까지 수집된 Kindle Store Category 데이터를 사용했다. 데이터에는 68,223명의 사용자, 61,934개의 상품과 982,619개의 리뷰로 구성되어 있다. 데이터에는 기본적인 사용자 정보, 상품 정보, 평점 및 리뷰 정보와 함께 다양한 추가 속성을 포함하고 있다. 데이터 속성에 대한 자세한 설명은 <표 8>의 예시와 같으며, 최소 5개 이상의 댓글을 작성한 사용자와 최소 5개 댓글을 받은 상품만 포함하고 있다.

데이터 희소성은 추천 시스템 연구에서 대표적으로 나타나는 문제점이다. 전자상거래 사이트에서 대부분의 상품은 인기도에 따라 선호도 정보가 편향되거나 상품에 대한 구매가 이루어지지 않아 선호도 데이터가 존재하지 않는 경우도 있다. 또한, 처음으로 구매하는 새로운 사용자의 경우에는 과거 구매 기록이나 선호도 데이터가 존재하지 않는다. 이와 같이 데이터의 희소성 문제는 사용자의 선호도 예측이 불가능하거나 추천 시스템의 성능을 측정하는데 한계가 존재한다. 따라서 본 연

〈표 8〉 데이터 속성 설명

Column Name	Description
asin	ID of the product, like B000FA64PK
helpful	helpfulness rating of the review - example: 2/3
overall	rating of the product
reviewText	text of the review(heading)
reviewTime	time of the review(raw)
reviewerID	ID of the reviewer, like A3SPTOKDG7WBLN
reviewerName	name of the reviewer
summary	summary of the review(description)
unixReviewTime	unix timestamp



〈그림 4〉 슬라이딩 윈도우를 이용한 검증

구에서는 데이터 희소성 문제를 극복하기 위해 기존 연구의 실험설계를 따라 구매기록을 많이 포함하고 있는 2013년 1월부터 2014년 7월까지 데이터를 선택하고 최소 10개 이하의 리뷰를 작성한 사용자와 최소 10개 이하의 리뷰를 받은 상품을 제외하고 사용했다. 본 연구에서 시간의 순서를 고려했을 때 추천 시스템의 성능을 평가하기 위해 슬라이딩 윈도우(Sliding Window) 평가방법을 채택했다. 예를 들어, 학습 데이터의 기간은 2013년 1월부터 2013년 9월까지 총 9개월을 사용하고, 검증 데이터의 기간은 2013년 10월부터 2013년 12월까지 총 3개월을 사용한다. 구체적인 구간 분할 방법은 <그림 4>와 같다.

3.4 성능평가

본 연구에서는 기존 추천 시스템 연구에서 상위 N개 추천 목록의 추천 성능을 평가할 때 널리 사용되고 있는 정밀도(Precision), 재현율(Recall)과 F1-Score 지표를 사용하여 제안한 추천 방법론의 성능을 측정하고자 한다. 정밀도는 모델이 True라고 분류한 것 중에서 실제 True인 것의 비율이고 재현율은 실제 True인 것 중에서 모델이 True라고 예측한 것의 비율이다. 정밀도와 재현율은 상충관계(Trade off)가 있으므로 본 연구에서는 식 (6)과 같이 F1-Score 지표를 사용하였다.

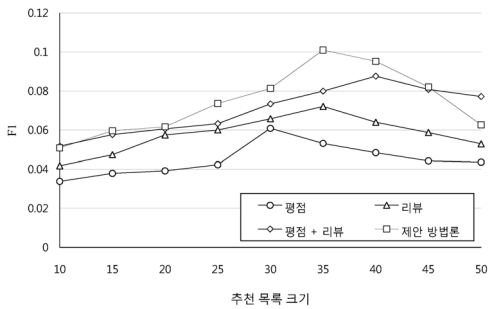
$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (6)$$

Ⅳ. 연구결과

본 연구에서는 개인화 추천 서비스에서 사용자의 선호도를 정교하게 예측하기 위해 평점과 리뷰를 순차적으로 고려하는 추천 방법론을 새롭게 제안했다. 본 연구에서 제안한 추천 방법론의 추천 성능을 측정하기 위해 평점만을 사용한 추천 방법론, 리뷰만을 사용한 추천 방법론과 평점과 리뷰

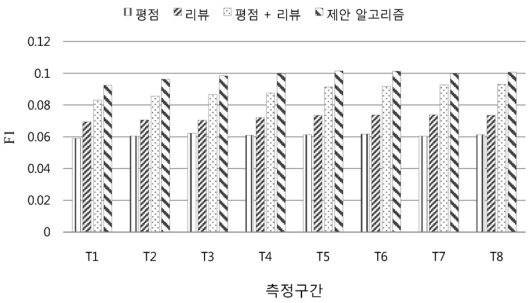
를 동시에 사용한 추천 방법론 비교하였다. 여기서 평정만을 사용한 방법론은 상품 간의 유사도를 계산하여 사용자에게 적합한 상품을 추천하는 CF 기반 추천 방법론이고, 리뷰만을 사용한 방법론은 상품 간의 리뷰 유사도를 계산하여 사용자에게 적합한 상품을 추천하는 TF-IDF 기반 추천 방법론이다. 평점과 리뷰를 동시에 사용하는 방법론은 평점과 리뷰의 유사도를 각각 CF, TF-IDF 기반으로 계산한 다음에 두개의 유사도를 통합하여 전체 사용자 간의 유사도를 산출하여 추천하는 방법이다.

다양한 크기의 추천 목록에서 추천 성능을 확인하기 위해 추천 목록의 크기를 10부터 50까지 설정하여 다양한 실험을 수행하였다. 구체적인 실험결과는 <그림 5>와 같다. 실험 결과에서 제안 방법론의 추천 목록의 크기가 10부터 45 사이에서는 기존 방법론보다 추천 성능이 높게 나타났다. 평점만을 사용한 방법론은 추천 목록의 크기가 30일 때 추천 성능이 가장 높게 나타냈다. 리뷰만을 사용하여 추천을 제공하는 경우 추천 목록의 크기가 35일 때 추천 성능이 가장 높고 평점과 리뷰의 유사도를 동시에 고려하여 추천을 제공하면 추천 목록의 크기가 40일 때 추천 성능이 가장 높게 나타냈다. 본 연구에서 제안한 방법론으로 추천을 제공할 경우 추천 목록의 크기가 35일 때 기존의 방법론 보다 높은 추천 성능을 나타내고 있다.



〈그림 5〉 추천 목록의 크기에따른 추천 정확도

여러 추천 방법론의 최적의 추천 목록의 크기를 적용하여 측정구간에 따른 추천 성능의 결과는 <그림 6>과 같다. 평점만을 사용하는 추천 방법론의 성능은 평균 0.06이고 리뷰만 이용하여 추천을 진행한 추천시스템의 F1은 약 0.07이다. 또한 평점과 리뷰의 유사도를 각각 산출한 뒤 두개의 유사도를 통합하여 추천을 진행한 추천시스템의 F1은 약 0.09이며, 본 연구에서 제안한 알고리즘인 먼저 평점 유사도를 이용하여 상품의 1차 이웃을 산출해낸 후 1차 이웃에 기반하여 리뷰데이터를 이용해 최종 이웃을 산출하여 추천을 진행한 추천 시스템의 F1은 약 0.10이다.

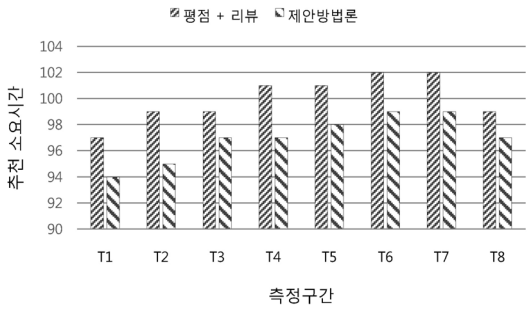


〈그림 6〉 측정기간에 따른 추천 정확도

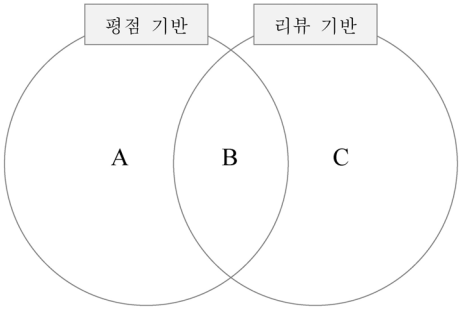
본 연구에서 추천 소요시간은 검증 데이터를 사용하여 추천을 제공할 때 소요되는 시간을 의미하며 추천 소요시간 값이 작을수록 추천 속도가 빠르다고 판단할 수 있다(Acilar and Arslan, 2009). 구체적인 실험결과는 <그림 7>과 같다. 기존의 평점과 리뷰를 동시에 고려하는 추천 방법론과 비교했을 때 본 연구에서 제안하는 방법론은 전체 측정구간에서 추천 소요시간이 더 우수한 성능을 보여주고 있다. 이는 기존 평점과 리뷰를 동시에 고려하는 연구의 한계점이라고 할 수 있는 확장성 문제점 개선에 기여할 수 있다.

본 연구에서 제안한 추천 방법론이 기존의 다른 추천 방법론보다 좋은 성능을 나타내는 이유는 유사도 계산을 통해 선정된 유사한 이웃 사용자의 범위가 <그림 8>과 같이 기존 추천 방법론과 다르기 때문이다. 기존의 평점만을 사용하는 추천 방법론에서 선정된 유사한 이웃 사용자의 범위는 A

와 B이고, 리뷰만을 사용하는 추천 방법론에서 선정된 유사한 이웃 사용자의 범위는 B와 C이다. 평점과 리뷰의 유사도를 동시에 고려하는 추천 방법론에서 선정된 이웃 사용자의 범위는 B와 A 또는 C의 일부분이다. 반면, 본 연구에서 제안한 추천 방법론에서 선정되는 유사한 이웃 사용자의 범위는 B이다. 이와 같이 동일한 선호도 데이터에서 선정되는 유사한 이웃 사용자의 범위가 작아지게 되면서 추천 성능은 기존 추천 방법론과 비교했을 때 높게 나타나고 추천 시스템이 연산하는 데이터의 개수가 감소되어 추천 소요시간도 줄어드는 장점이 있다.



〈그림 7〉 측정기간에 따른 추천 추천 소요시간



〈그림 8〉 유사한 이웃 사용자 선정

또한 정확도 비교를 통해 다음을 확인하였다. 첫째, 데이터에서 평점도 선호를 반영하고 리뷰도 선호를 반영한다. 그러나 리뷰를 이용한 추천 결과가 평점을 이용한 추천 결과보다 정확도가 높다. 이렇게 나온 원인은 본 논문에서 사용한 데이터가 온라인 서적이기 때문이다. 즉, 온라인 서적의 특성상 사용자들이 리뷰를 남길 때 서적에 대한 설명을 자세하게 작성하고 서적에 대한 사용자의 감정이 잘 나타나기에 리뷰 키워드의 특징이 잘 반영되기 때문이다. 둘째, 평점이나 리뷰만을 이용하여 추천을 진행하는 경우보다 두가지 데이터를 동시에 사용하였을 경우에 추천 정확도가 더 높다. 이는 평점과 리뷰를 동시에 고려하는 것이 사용자의 선호에 대한 정보가 더 상세하기 때문이다. 셋째, 평점 유사도와 리뷰 유사도를 각각 따로 고려한 뒤 통합하여 추천을 진행하는 경우보다 순차적으로 고려하는 경우에 정확도가 더 높다. 이러한 결과가 나타나는 원인은 본 논문에서 제안한 알고리즘은 평점과 리뷰를 순차적으로 이용하여 추천을 진행함으로써 평점과 리뷰의 유사도를 각각 산출한 뒤 두 개의 유사도를 통합하여 추천을 진행하는 경우보다 평점과 리뷰의 특성을 잘 반영하였기 때문이다.

Ⅴ. 결 론

본 연구에서는 사용자의 명시적 평점(Explicit Rating)과 구매 여부 등을 나타내는 암묵적 데이터(Implicit Data)를 사용하여 개인화 서비스를 제공하는 기존 추천 시스템 연구의 한계를 개선하기 위해 새로운 추천 방법론은 제안했다. 제안한 추천 방법론의 성능을 검증하기 위해 전자상거래 사이트에서 사용자가 특정 상품을 구매할 때 중요하게 생각하는 평점과 리뷰 데이터를 사용하였다. 또한, 평점과 유사도를 각각 계산하면 확장성 문제점을 나타낼 수 있는 기존 연구의 한계를 개선하기 위해 평점과 유사도를 순차적으로 계산하여 추천 소요시간을 단축시켰다. 제안한 추천 방법론을 통해 사용자가 특정 상품에 작성한 리뷰에서 키워드를 추출하고 TF-IDF 기법을 적용하여 유사도를 계산하여 이를 추천 시스템에 반영했다. 실험 결과, 기존 추천 방법론보다 본 연구에서 제안한 추천 방법론이 더

우수한 추천 성능이 나타내고 있으며,기존 연구에서 나타냈던 확장성 문제점도 크게 개선되었다.

본 연구의 시사점은 다음과 같다. 첫째, 전자상거래 사이트에서 리뷰 데이터는 사용자가 특정 상품을 구매할 때 중요하게 생각하는 부분이다. 따라서 리뷰 데이터를 정교하게 분석하여 사용자에게 맞춤형 추천 서비스를 제공하는 것이 중요한 작업이라고 할 수 있다. 기존 추천 시스템의 연구는 주로 정량적인 평점 데이터를 사용하여 추천 시스템을 구축하고 사용자의 선호도를 예측했다. 본 연구에서는 전자상거래 사이트에 사용자가 직접 작성한 리뷰 데이터에 TF-IDF 기법을 적용을 통해 선호도를 정교하게 분석하여 추천 시스템에 반영했다. 특히 기존 리뷰 데이터 관련 연구는 주로 감성분석 기법에 국한되었다. 본 연구에서는 리뷰 데이터에서 선호도를 정교하여 분석하여 추천 시스템에 반영하여 사용자에게 맞춤형 추천 서비스를 제공하는 새로운 추천 방법론을 제안했다. 본 연구는 주로 평점 데이터만을 사용하는 기존 추천 시스템 연구에 리뷰 데이터를 추가로 적용함으로써 추천 시스템 연구의 확장에 기여하고 있다. 둘째, 본 연구에서 제안한 추천 방법론은 평점과 리뷰 데이터를 동시에 수집하는 다른 분야에서도 사용할 수 있다. 최근에 많은 전자상거래 사이트에서사용자가 직접 상품에 대한 리뷰를 작성하는 서비스 제공을 시작하고 있다. 본 연구에서는 전자상거래 분야에만 제안 방법론을 적용하여 추천 서비스를 제공하였으나, 평점과 리뷰 데이터를 동시에 수집하는 다른 분야에도 사용할 수 있다. 이를 통해 리뷰 데이터의 키워드를 추출하여 추천 시스템에 반영하고 사용자의 선호도를 정교하게 예측할 수 있다.

본 연구의 한계점과 향후 연구 계획은 다음과 같다. 첫째, 본 연구에서는 먼저 사용자의 평점만을 사용하여 유사한 이웃 사용자를 산출하고 다음 리뷰 데이터를 사용하여 최종 유사한 이웃 사용자를 산출했다. 향후 연구에서는 먼저 리뷰 데이터만을 사용하여 유사한 이웃 사용자를 산출하고 다음으로 평점을 사용하여 최종 유사한 이웃 사용자를 산출하였을 때 추천 성능에 대하여 확인하고자 한다. 또한 다른 상품에 대하여 제안 추천 방법론을 적용하였을 때 추천 성능에 대하여 검증하고자 한다. 예를 들어 화장품, 의류 혹은 상품 분야에 적용했을 때 평점과 리뷰가 사용자의 구매 여부에 어떤 영향을 미치는지 확인하는 것도 고려해 볼 수 있다. 또한, 추천 알고리즘을 딥러닝 기법으로 구현하였을 때의 성능 개선에 어떻게 반영되는지도 향후 연구 과제로 남아있다.

참 고 문 헌

[1] 김병만, 이경, 김시관, 임은기, 김주연, “추천 시스템을 위한 내용기반 필터링과 협력필터링의 새로운 결합 기법”, *정보과학회논문지: 소프트웨어 및 응용*, 제31권, 제3호, 2004, pp. 332-342.

[2] 박대서, 김화종, “TF-IDF 기반 키워드 추출에서의 의미적 요소 반영을 위한 결합벡터 제안”, *한국정보기술학회논문지*, 제16권, 제2호, 2018, pp. 1-16.

[3] 박종학, 조윤호, 김재경, “사회연결망: 신규고객 추천문제의 새로운 접근법”, *지능정보연구*, 제15권, 제1호, 2009, pp. 123-140.

[4] 연다인, 박가연, 김희웅, “텍스트 마이닝 기반 사용자 경험 분석 및 관리: 스마트 스피커 사례”, *Information Systems Review*, 제22권, 제2호, 2020, pp. 77-99.

[5] 유은순, 최건희, 김승훈, “TF-IDF와 소설 텍스트의 구조를 이용한 주제어 추출 연구”, *한국컴퓨터정보학회논문지*, 제20권, 제2호, 2015, pp. 121-129.

[6] 이상기, 이병섭, 박병용, 황혜경, “나이브베이즈 분류모델과 협업필터링 기반 지능형 학술논문 추천시스템 연구”, *정보관리연구*, 제41권, 제4호, 2010, pp. 227-249.

[7] 이성직, 김한준, “TF-IDF의 변형을 이용한 전

자뉴스에서의 키워드 추출 기법", *한국전자거래학회지*, 제14권, 제4호, 2009, pp. 59-73.

[8] 전병국, 안현철, "사용자 리뷰 마이닝을 결합한 협업 필터링 시스템: 스마트폰 앱 추천에의 응용", *지능정보연구*, 제21권, 제2호, 2015, pp. 1-18.

[9] 현지연, 유상이, 이상용, "평점과 리뷰 텍스트 감성분석을 결합한 추천시스템 향상 방안 연구", *지능정보연구*, 제25권, 제1호, 2019, pp. 219-239.

[10] Acilar, A. M. and A. Arslan, "A collaborative filtering method based on artificial immune network", *Expert Systems with Applications*, Vol.36, No.4, 2009, pp. 8324-8332.

[11] Batmaz, Z., A. Yurekli, A. Bilge, and C. Kaleli, "A review on deep learning for recommender systems: Challenges and remedies", *Artificial Intelligence Review*, Vol.52, No.1, 2019, pp. 1-37.

[12] Bennett, J. and S. Lanning, "The Netflix Prize", In *Proceedings of KDD Cup and Workshop*, Vol.2007, 2007, pp. 3-6.

[13] Berry, M. W. and M. Castellanos, *Survey of Text Mining*, Springer-Verlag, New York, NY, 2004.

[14] Bobadilla, J., F. Ortega, A. Hernando, and A. Gutiérrez, "Recommender systems survey", *Knowledge-Based Systems*, Vol.46, 2013, pp. 109-132.

[15] Cho, Y. H. and J. K. Kim, "Application of web usage mining and product taxonomy to collaborative recommendations in e-commerce", *Expert Systems with Applications*, Vol.26, No.2, 2004, pp. 233-246.

[16] Choi, I. Y., H. S. Moon, and J. K. Kim, "Accessing personalized recommendation services using expectancy disconfirmation theory", *Asia Pacific Journal of Information Systems*, Vol.29, No.2, 2019a, pp. 203-216.

[17] Choi, I. Y., Y. U. Ryu, and J. K. Kim, "A recommender system based on personal constraints for smart tourism city", *Asia Pacific Journal of Tourism Research*, 2019b, pp. 1-14.

[18] Claypool, M., A. Gokhale, T. Miranda, P. Murnikov, D. Netes, and M. Sartin, "Combing content-based and collaborative filters in an online newspaper", In *Proceedings of Workshop on Recommender Systems: Algorithms and Evaluation*, 1999.

[19] Covington, P., J. Adams, and E. Sargin, "Deep neural networks for youtube recommendations", In *Proceedings of the 10th ACM Conference on Recommender Systems*, 2016, pp. 191-198.

[20] Goldberg, D., D. Nichols, B. M. Oki, and D. Terry, "Using collaborative filtering to weave an information tapestr", *Communications of the ACM*, Vol.35, No.12, 1992, pp. 61-71.

[21] Gupta, V. and G. S. Lehal, "A survey of text mining techniques and applications", *Journal of Emerging Technologies in Web Intelligence*, Vol.1, No.1, 2009, pp. 60-76.

[22] Herlocker, J. L., J. A. Konstan, and J. Riedl, "Explaining collaborative filtering recommendations", In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, 2000, pp. 241-250.

[23] Herlocker, J. L., J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender systems", *ACM Transactions on Information Systems(TOIS)*, Vol.22, No.1, 2004, pp. 5-53.

[24] Jannach, D., M. Zanker, A. Felfernig, and G. Friedrich, *Recommender Systems: An Introduction*, Cambridge University Press, 2010.

[25] Kim, H. K., H. Y. Oh, J. C. Gu, and J. K. Kim, "Commanders: A recommendation procedure for online book communities", *Electronic Commerce Research and Applications*, Vol.10, No.5, 2011a,

pp. 501-509.

[26] Kim, H. K., Y. U. Ryu, Y. Cho, and J. K. Kim, "Customer-driven content recommendation over a network of customers", *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, Vol.42, No.1, 2011b, pp. 48-56.

[27] Kim, J. K., H. K. Kim, H. Y. Oh, and Y. U. Ryu, "A group recommendation system for online communities", *International Journal of Information Management*, Vol.30, No.3, 2010, pp. 212-219.

[28] Kim, J. K., H. S. Moon, B. J. An, and I. Y. Choi, "A grocery recommendation for off-line shoppers", *Online Information Review*, Vol.42, No.4, 2018, pp. 468-481.

[29] Kim, J. K., Y. H. Cho, S. T. Kim, and H. K. Kim, "A personalized recommender system for mobile commerce applications", *Asia Pacific Journal of Information Systems*, Vol.15, No.3, 2005, pp. 223-241.

[30] Lee, H. I., I. Y. Choi, H. S. Moon, and J. K. Kim, "A multi-period product recommender system in online food market based on recurrent neural networks", *Sustainability*, Vol.12, No.3, 2020, p. 969.

[31] Li, L. and R. Zhang, "Recommended study of the flow of information based on TF-IDF", *International Journal of Hybrid Information Technology*, Vol.8, No.8, 2015, pp. 191-200.

[32] Linden, G., B. Smith, and J. York, "Amazon.com recommendation: Item-to-item collaborative filtering", *IEEE Internet Computing*, Vol.7, No.1, 2003, pp. 76-80.

[33] Park, Y., S. Park, W. Jung, and S. G. Lee, "Reversed CF: A fast collaborative filtering algorithm using a k-nearest neighbor graph", *Expert Systems with Applications*, Vol.42, No.8, 2015, pp. 4022-4028.

[34] Ricci, F., L. Rokach, and B. Shapira, *Introduction to Recommender Systems Handbook*, In *Recommender Systems Handbook* (pp. 1-35), Springer, Boston, MA, 2011.

[35] Schafer, J. B., D. Frankowski, J. Herlocker, and S. Sen, "Collaborative filtering recommender systems", In *The Adaptive Web*, Springer, 2007, pp. 291-324.

[36] Shani, G. and A. Gunaward, *Evaluating Recommendation Systems*, In *Recommender Systems Handbook* (pp. 257-297), Springer, Boston, MA, 2011.

[37] Su, X. and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques", *Advances in Artificial Intelligence*, Vol.2009, 2009, pp. 1-19.

[38] Zhang, Z., D. Zhang, and J. Lai, "urCF: User review enhanced collaborative filtering", In *Proceedings of the 20th Americas Conference on Information Systems*, 2014.

[39] Zheng, L., V. Noroozi, and P. S. Yu, "Joint deep modeling of users and items using reviews for recommendation", In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, 2017, pp. 425-434.

Information Systems Review
Volume 23 Number 1
February 2021

Development of a Hybrid Recommender System Using Review Data Mining: Amazon Kindle Store Data Analysis Case

Yihua Zhang* · Qinglong Li** · Ilyoung Choi*** · Jaekyeong Kim****

Abstract

With the recent increase in online product purchases, a recommender system that recommends products considering users' preferences has still been studied. The recommender system provides personalized product recommendation services to users. Collaborative Filtering (CF) using user ratings on products is one of the most widely used recommendation algorithms. During CF, the item-based method identifies the user's product by using ratings left on the product purchased by the user and obtains the similarity between the purchased product and the unpurchased product. CF takes a lot of time to calculate the similarity between products. In particular, it takes more time when using text-based big data such as review data of Amazon store. This paper suggests a hybrid recommendation system using a 2-phase methodology and text data mining to calculate the similarity between products easily and quickly. To this end, we collected about 980,000 online consumer ratings and review data from the online commerce store, Amazon Kinder Store. As a result of several experiments, it was confirmed that the suggested hybrid recommendation system reflecting the user's rating and review data has resulted in similar recommendation time, but higher accuracy compared to the CF-based benchmark recommender systems. Therefore, the suggested system is expected to increase the user's satisfaction and increase its sales.

Keywords: Review data mining, Text Mining, Recommendation System, Collaborative Filtering

* Master Student, Department of Big Data Analytics, KyungHee University
** Master Student, Department of Big Data Analytics, KyungHee University
*** Lecturer, Graduate School of Business Administration & AI Management Research Center, KyungHee University
**** Corresponding Author, Professor, School of Management & Department of Big Data Analytics, KyungHee University

◐ 저 자 소 개 ◑

장 예 화 (yihua0124@khu.ac.kr)

중국 연변대학교 통신학과를 졸업하고, 현재 경희대학교 일반대학원 빅데이터응용학과 석사과정에 재학 중이며, 주요 관심분야는 추천시스템, 딥러닝, 텍스트마이닝, 데이터 분석 등이다.

이 청 용 (leecy@khu.ac.kr)

경희대학교에서 경영학 학사를 취득하고, 현재 동 대학원 빅데이터응용학과 석사과정에 재학 중이다. 주요 관심 분야는 데이터마이닝, 딥러닝, 추천 시스템, 자연어 처리, 빅데이터시각화 등이다.

최 일 영 (choice102@khu.ac.kr)

경희대학교에서 경제학 학사, 동 대학원에서 경영정보시스템 전공으로 경영학 석사, 박사 학위를 취득하였다. 주요 관심분야는 빅데이터 분석, 딥러닝, 추천 시스템, 그린 비즈니스/IT, 비즈니스 인텔리전스, 사회네트워크분석 등이며 Information Technology and Management, International Journal of Information Management, Online Information Review, 경영과학회지, 경영과학, 정보관리학회지, 지능정보연구 등 다수의 학술지에 논문을 게재하였다.

김 재 경 (jaek@khu.ac.kr)

서울대학교에서 산업공학학사, 한국과학기술원에서 경영정보시스템 전공으 로 석사 및 박사학위를 취득하였으며 현재 경희대학교 경영대학 및 빅데이터응용학과 교수로 재직하고 있다. 미국 미네소타 주립대학교 그리고 텍사스 주립대학교 (달라스)에서 교환교수를 역임하였다. 주요 관심 분야로는 개인화서비스, 추천시스템, 빅데이터, 및 딥러닝 등이다. IEEE Transaction on Services Computing, IEEE Transaction on SMC-A, International Journal of Human Computer Studies, International Journal of Information Management, Information and Management, Expert Systems with Applications, Applied Artificial Intlligence, 등 다수의 학술지에 논문을 게재하였다. 현재 4단계 BK21사업 연구단장 (빅데이터분야) 및 AI 비즈니스 연구센터 센터장을 맡고 있다.

논문접수일 : 2020년 09월 15일 **게재확정일 : 2020년 12월 29일**
1차 수정일 : 2020년 12월 28일