

영상 콘텐츠 시청 데이터를 활용한 개인 맞춤형 도서 추천 시스템

임예빈¹, 이경민¹, 김유진², 이서영¹, 김현희¹

¹동덕여자대학교 정보통계학과

²동덕여자대학교 경영학과

20191068@dongduk.ac.kr, 20191058@dongduk.ac.kr, 20190287@dongduk.ac.kr,

20201064@dongduk.ac.kr, heekim@dongduk.ac.kr

Personalized book recommendation system using video content viewing data

Yea Bin Lim¹, Gyeong Min Lee¹, Yu Jin Kim², Seo Young Lee¹,

Hyon Hee Kim¹

¹Dept. of Statistics and Information Science, Dongduk Women's University

²Dept. of Business Administration, Dongduk Women's University

요 약

최근 성인 독서량은 지속적으로 감소하는데 비해 영상 콘텐츠 소비가 증가하고 있다. 이에 따라 새로운 사용자에게 대한 선호도 및 행동 패턴에 대한 정보가 없고 새로운 도서에 대한 사용자 평가나 구매 정보가 부족해 콜드 스타트 문제와 데이터 희소성 문제가 발생하고 있다. 본 논문에서는 영상물 콘텐츠 기반 도서 하이브리드 추천 시스템을 제안하였다. 제안하는 추천 시스템은 영상물의 콘텐츠를 활용하여 콜드 스타트 문제와 데이터 희소성 문제를 해결할 수 있을 뿐만 아니라, 전통적인 도서 추천 시스템에 비해 성능이 향상되었고 장르, 줄거리, 평점 정보 기반 사용자 취향 정보까지 모두 반영된 질 높은 추천 결과까지 확인할 수 있었다.

1. 서론

2021년 통계청에 따르면, 성인 중 일반 도서를 1년에 한 권 이상 읽은 성인의 비율은 46.9%에 불과하며 독서 장애 요인으로 콘텐츠 이용을 가장 많이 답변했다[1]. 성인 독서량은 감소하는 반면 영상 콘텐츠 소비가 증가함에 따라 새로운 사용자의 선호도 및 행동 패턴 정보와 도서 소비와 관련한 데이터가 점점 부족하여 콜드 스타트 (Cold-Start) 문제와 데이터 희소성 (Data Sparsity) 문제가 도서 추천 시스템에서 점차 증가하고 있다[2].

이러한 문제를 해결하기 위해 본 논문에서는 영상 콘텐츠 소비 패턴을 활용하여 도서를 추천하는 교차도메인 추천 시스템을 설계 및 개발하였다. 먼저, 콘텐츠를 활용하여 추천하는 콘텐츠 기반 필터링 추천 시스템과 비슷한 영상 콘텐츠 취향을 가진 사용자들의 도서 정보들을 모아 추천하는 협업 필터링 추천 시스템을 구현하였다. 그리고 두 개의 추천 시스템에서 가중치를 조절하여 콘텐츠별 장르, 줄거리, 평점 정보 모두를 적절히 반영한 하이브리드 추천 시스템을 개발하였다.

본 연구에서 제안하는 교차 도메인에서의 하이브리드 도서 추천 시스템은 단순히 도서 정보만을 활용하는 도서 추천 시스템보다 성능이 우수할 뿐만

아니라 영화 시청 기록과 콘텐츠를 활용함으로써, 데이터 희소성과 콜드 스타트 문제를 해결하여 개인 맞춤형 도서 추천이 가능하도록 하였다. 향후 성인 독서 장애 요인인 독서 습관 부재를 해결하고 도서관 및 온·오프라인 서점에 기술을 접목하여 도서 산업의 부흥에 한층 기여할 것으로 기대된다.

2. 데이터 소개

본 연구에서는 사용자의 영상 콘텐츠 취향을 기반으로 도서를 추천하기 위해 교보문고의 베스트셀러 도서 100권과 역대 박스오피스 상위 100위 영화 데이터를 활용했다. 먼저 영상 콘텐츠와 도서에 대해 사용자 평점을 제공하는 왓차피디아에서 평점을 남긴 최신순으로 사용자를 정렬했다. 그 후, 해당 사용자가 평점을 남긴 영화와 도서를 추가로 크롤링 하여 총 624개의 영상 콘텐츠, 716개의 도서, 2251개의 이용자 평점 데이터를 수집했다.

수집한 도서와 영화에서 장르, 줄거리, 사용자 평점을 변수로 활용했다. 장르는 인문, 자기 계발, 시/에세이, 소설, 과학, 역사/문화, 예술/대중문화 7개로 재분류하였다. 줄거리는 명사를 추출하여 토큰화를 진행한 뒤 관련 없는 단어를 불용어 리스트로 설정해 변수의 특성을 높였다. 또한 사용자의 평점이 없

는 콘텐츠의 경우, 딥러닝을 활용해 평점을 예측하여 결측치를 채웠다.

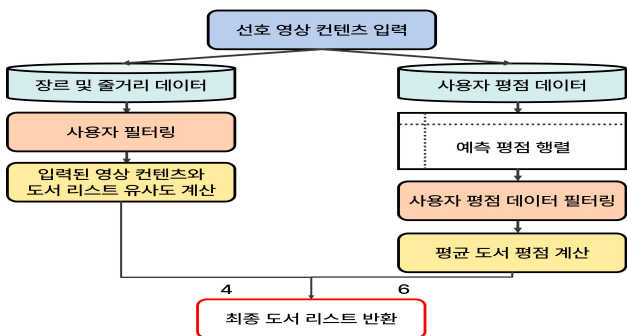
3. 교차 도메인에서의 하이브리드 추천 모델

본 연구는 콘텐츠 기반 필터링 추천 시스템과 협업 필터링 기반 추천 시스템을 구현한 후 가중치를 조절하여 동시에 반영한 하이브리드 추천 시스템을 구현했다. 예측 평점과 줄거리 유사도의 반영 비율을 달리하며 상위 5개의 도서를 반환하도록 했다. 이때, 가중치를 조정해 콘텐츠 별 장르, 줄거리, 유사한 취향을 가진 사용자들의 평점 정보 모두를 적절하게 반영하도록 했다.

하이브리드 추천 시스템의 예측 성능을 최대화하기 위해 콘텐츠 기반 필터링과 협업 필터링의 적절한 가중치를 찾는 과정을 수행하였다. 사용한 수식은 다음과 같다 :

$$(\lambda_1 * \text{유사도}) + (\lambda_2 * \text{평점}) \tag{1}$$

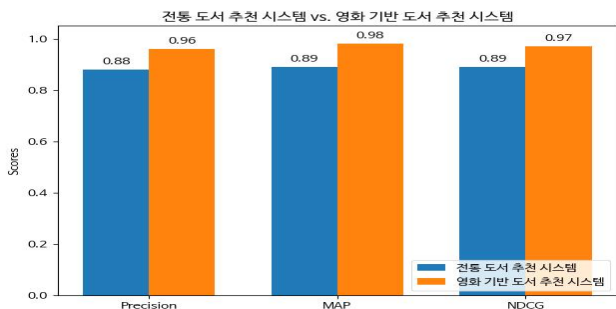
여러 실험 결과, 4:6로 반영했을 때 가장 성능이 높았다. 그림 1은 제안하는 하이브리드 추천 시스템의 도식도이다.



(그림 1) 하이브리드 추천 시스템 알고리즘

4. 성능 평가

본 연구에서는 추천 시스템의 성능을 추천하기 위해 Precision, MAP(Mean Average Precision), Normalized Discounted Cumulative Gain (NDCG) 지표를 사용하였다.



(그림 2) 추천 시스템의 성능 비교

그림 2는 본 논문에서 제안하는 영상 콘텐츠 도서 추천 시스템과 도서 데이터만을 가지고 추천한 전통 도서 추천 시스템의 성능을 비교한 결과이다. Precision, MAP, NDCG 세 가지 측정 지표의 경우 모두 제안하는 하이브리드 크로스 도메인 추천 시스템이 높은 성능을 보였음을 알 수 있다.

다음으로 3가지 영상 콘텐츠 기반 도서 추천 시스템에 영상 콘텐츠를 입력했을 때의 도서 추천 결과를 비교했다.

표 1은 여러 영상을 입력했을 때의 결과이다. 예를 들어, ‘광해’와 ‘봉오동 전투’를 입력했을 때, 콘텐츠 기반 필터링에선 단순히 내용이 비슷한 책으로 추천이 이뤄졌다. 그러나 하이브리드 추천에서는 줄거리와 사용자 평점까지 포함하여 추천이 이뤄졌다. 단순 키워드 매칭을 보면 ‘최소한의 한국사’가 ‘나는 소망한다 내게 금지된 것을’ 보다 더욱 일치하지만, 하이브리드 추천 시스템을 통해 입력한 영상 콘텐츠를 선호한 사용자들이 높게 평가한 책인 ‘나는 소망한다 내게 금지된 것을’이 ‘최소한의 한국사’보다 더 높은 순위로 추천된 것을 확인할 수 있다.

<표 1> 영상 콘텐츠 기반 도서 추천 시스템 결과

순위	콘텐츠 기반 추천	하이브리드 추천	순위 변동
1	역사의 쓸모(3.9)	역사의 쓸모(3.9)	-
2	슬기로운 세계사(3.4)	슬기로운 세계사	-
3	최소한의 한국사(3.8)	나는 소망한다 내게 금지된 것을(4.1)	5->3
4	30개 도서로 읽는 한국사(3.3)	최소한의 한국사(3.8)	3->4

5. 결론

본 연구는 도서를 통한 도서 추천 시스템의 한계를 탈피하고자 사용자의 영상 콘텐츠 취향을 기반으로 도서를 추천하는 모델을 제안하였다. 기존의 전통적인 도서 추천 시스템의 경우, 쿨드 스타트 문제가 발생한다는 시대적인 단점이 존재하는 반면, 영상물 기반으로 도서를 추천하는 하이브리드 추천 시스템을 활용하였을 경우, 새로운 도메인이 추가되었음에도 기존 도서 추천 시스템과 견주어 보았을 때, 예측 성능이 향상되었고 줄거리, 장르, 평점 정보 모두 반영된 질 높은 추천 결과까지 확인할 수 있었다.

참고문헌

[1] 문가영, “책 덮은 대한민국,, 당신은?... 1년간 단 한권도 안읽은 성인 53%”, 매일경제, 2022, <https://www.mk.co.kr/news/society/10482768>
 [2] 통계청(2022), 「국민독서실태조사 통계정보보고서」, 12-14