메타 가중치 학습을 활용한 내용 기반의 맞춤형 영화 추천시스템 설계 및 구현

안현우, 유해운, 김대열
티비스톰
grabro@tvstorm.com, syrics@tvstorm.com, wagoon0004@tvstorm.com

Design and Implementation of Contents-based Customized movie recommendation system using meta weight learning

Hyeon Woo An, Hea Woon You, Dea Yeol Kim
TVSTORM

요 약

최근, 디지털 콘텐츠 산업이 폭발적으로 성장됨에 따라 고객 유치를 위한 개인화 추천 기술들이 많은 주목을 받고 있다. 개인화 추천 방식들을 큰 갈래로 나누어 본다면 협업 필터링 기술과 내용 기반 기술로 나눌 수 있다. 협업 필터링의 경우 개인화 추천에는 적합하지만 사용자 평가 데이터의 양이 방대해야 하며 초기에 평가자가 없는 콘텐츠에 대해 추천할 수 없는 초기 평가자 문제가 존재한다. 따라서 매일 방대한 양의 콘텐츠가 편입되는 분야에서 사용하기에 큰 결점이 될 수 있다. 본 논문에서는 영화들의 정보가 담긴 데이터 셋과 사용자 평가 데이터, 그리고 사용자의 선호 기준을 의미하는 메타 가중치를 활용한 내용 기반의 맞춤형 영화 추천 시스템을 제안한다. 논문에서는 먼저, 영화를 고를 때 일반적으로 중요시 보는 속성들을 활용하여 영화의 특징 벡터를 구성하고, 이를 사용자 평가와 결합하여 개인의 선호에 대한 특징 벡터를 구성하는 방법을 제안하며, 구성된 데이터와 코사인 유사도, 메타 가중치를 활용하여 사용자 선호와 유사한 영화들을 도출하는 방법을 제안한다. 또한, 평가데이터를 활용하여 구현된 추천시스템의 검증 프로세스를 구성하고, 검증 프로세스를 활용한 손실 함수를 설계하여 적합한 메타 가중치를 학습하는 방법을 제시한다. 본 논문에서 제안하는 시스템은 다수의 속성을 조합하여 활용하므로 추천 결과가 과도하게 특수화 되지 않을 수 있으며, 메타 가중치라는 요소를 통해 더욱 개인화 된 추천을 제공할 수 있다.

1. 서론

근래에 들어 추천시스템은 초개인화라는 트랜드에 맞춰 발빠르게 변화하고 있다. 콘텐츠가 범람하는 환경에서 사용자의 불편함은 해소시키고 소비를 증대시킨다는 목적을 넘어 본인도 자각하지 못한 취향이나 기분, 날씨, 특일 등의 외부 환경까지 고려한 추천을 제공하여 이용자의 플랫폼 만족도를 높인다는 것이다. 이는 비단 콘텐츠 소비의 용도뿐 아니라 여러 분야에도 활용될 수 있는데 S 카드사의 시계열 소비 패턴 분석을 통한 맞춤형 쿠폰 제공이 그 예이다. 사용자의 취향에 따라 소비의 방향이 천차만별인 OTT 서비스에서도 이는 매우 중요한 과제이며 최근 동향으로 비추어 본다면 고객 유치와 밀접한 관련이 있다고 말할 수 있다.

일반적으로 개인화 추천시스템은 사용자 개개인의 선호를 분석하고 추천하는 형태로 이루어지는데, 주로 협업 필터링과 내용기반 필터링 방법이 사용된다.

협업 필터링의 경우 사용자들의 평가정보를 토대로 콘텐츠의

잠재 정보를 추출하거나 유사한 사용자의 평가를 반영하여 직접적인 접촉이 없는 콘텐츠의 추천을 가능하게 해준다. 하지만 평가 이력이 얼마 없는 사용자에 대한 추천이 어려운 콜드 스타트 문제(Cold-Start problem)나 아무도 평가하지 않은 콘텐츠에 대한 추천이 불가능한 초기 평가자 문제(First-Rater problem) 등으로 인해 실 적용에 있어 많은 애로사항이 존재한다.

내용기반 필터링의 경우에는 콘텐츠가 자체적으로 갖고 있는 속성들을 분석하여 콘텐츠를 구분하고 사용자의 선호와 연결하여 유사한 콘텐츠를 추천해주는 방식이다. 협업 필터링과 비교하였을 때 사용자 선호의 반영이 콘텐츠 분석과 독립되어있기 때문에 선호를 이루는 특징에 대한 추가나 개선이 용이하며, 선호를 판단하는 데 있어 평가 기반이 아닌 프로파일이나 사용자 관련 정보 등을 기반으로 할 수 있어 콜드 스타트나 초기 평가자 문제에 비교적 자유로운 편이다. 하지만 추천의 결과에 있어 그 다양성이 상대적으로 제한되는 과도한 특수화 문제(Over-specailze problem)와 서로 다른 종류의 콘텐츠에 대한 추천이 힘들다는 단점이 존재한다[1].

본 논문은 이러한 추천시스템의 동향과 각 기법의 장단점에 착안하여 내용 기반의 필터링을 통해 사용자 개인이 갖고 있는 선호의 기준을 깊이 있게 파악하고 시즌이나 특일, 기상 등의 영향요인을 반영할 수 있는 영화 추천 시스템을 제안한다.

2. 본론

본 연구에서는 내용 기반의 피처 추출을 위한 영화 정보 참조를 위해 IMDB 라는 영화 플랫폼에서 제공하는 데이터 셋을 활용하였으며, 사용자 선호 파악과 메타 가중치 학습에 대한 실험을 진행하기 위해 IMDB 와의 결합 속성을 갖고 있는 Movielens 데이터를 활용하였다[2].

IMDB 데이터는 영화에 대한 많은 속성 정보를 포함하고 있다. 본 연구에서는 다양한 영화의 속성들 중 사용자의 선호, 또는 선정 기준에 영향을 미친다고 판단된 줄거리, 장르, 감독, 배우, 인지도 5 개의 속성들을 영화를 설명하는 메타 데이터로 활용하였다. 또한 사용자의 선호를 파악하기 위해 선호도가 반영된 평가 데이터 혹은 리뷰 데이터와 함께 개인화 된 메타 가중치를 활용한다.

제안하는 추천 시스템은 그림 1 과 같이 동작한다.

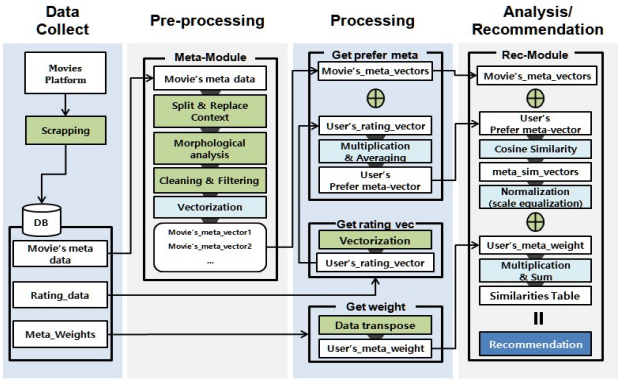

그림 1. 메타가중치를 적용한 개인화 추천시스템의 개요

추천 시스템은 크게 상기한 데이터를 수집하는 과정과, 영화 메타 데이터를 벡터화하기 위한 전처리 과정, 사용자의 선호 벡터를 생성하기 위한 가공 과정, 위 두 결과 벡터를 통해 추천 결과를 도출하는 추천 과정으로 이루어진다.

전처리 과정에서는 기본적인 문장 분석과 정제 과정을 포함하며, TF-IDF 를 통한 줄거리 벡터화와, 장르, 감독, 배우의 경우 참/거짓 벡터화, 인지도의 경우 투표수/개봉기간의 순위에 따른 수치화가 적용된다. 이때, 줄거리의 경우 사람 이름을 그대로 포함하여 반영할 경우, TF-IDF 에 의해 사람 이름을 강하게 특징으로 반영되는 현상이 발생한다. 본 연구에서는 이를 제거하기 위해 구글에서 개발한 BERT 를 활용하여 사람이름에 대한 제거 작업을 진행하였다[3].

가공 과정부터는 추천의 런타임 과정에 포함된다. 즉 추천에 대한 요청이 진행 되는 순간부터가 가공 과정이라 할 수 있으며 전처리 과정에서 도출된 영화의 특징 벡터를 토대로 사용자의 선호 벡터를 만드는 작업이 포함된다. 이때, 사용자 선호 벡터 U 는 사용자가 평가한 특정 n 개의 영화들에 대해 선호 점수로 가중치화 된 평가데이터 집합 R 과 인지도를 제외한 영화 특징 벡터 V 를 토대로 아래 수식을 통해 생성한다. 여기서 R 이 가리키는 영화의 인덱스 집합은 M 으로 표현한다.

$$U = \prod_{j=1}^{4}\left[\frac{1}{n}\sum_{i=1}^{n} v_{j,m_i}\, r_i\right] \quad (1)$$

인지도의 경우 나머지 4 개의 메타 데이터와 다르게, 영화의 품질을 대변하는 유일한 메타 데이터이기 때문에 가중치가 포함되지 않는 $V_{5,M}$의 단순 평균으로 계산한다.

가공 과정에서 일반적으로 U 는 벡터의 크기에 비해 값이 희박한 희소 행렬(sparse matrix)형태를 띄우고 있는데, 압축된 형태로 관리하여 벡터가 가진 실질적인 크기에 비해 빠른 처리를 유도하는 것이 주된 목표라 할 수 있다.

추천 과정에서는 사용자 선호 벡터 U 와 영화의 특징 벡터 V

간의 코사인 유사도 테이블을 구하고, 정규화 과정을 통해 각 메타 벡터의 유사도가 균등하게 반영되도록 적용한 뒤, 개인의 메타 가중치를 적용하여 개인화 된 추천 결과를 제공하는 작업을 진행한다. 결과적으로 사용자 선호 벡터는 콘텐츠 속성에 대한 선호로, 메타 가중치는 각 속성에 대한 선호로 반영된다. 즉 메타 가중치는 개인마다 상이한 선호의 기준을 반영하는데 사용되며 별도의 학습 과정을 통해 학습된다.

본 연구에서는 일반적으로 추천시스템의 성능을 평가하는 CTR(Click Through Rate)과 유사하게 아래 그림 2 와 같은 검증 프로세스를 제안한다.

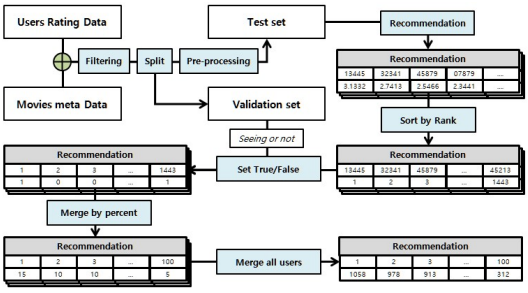

그림 2. 검증 과정

검증 과정을 요약하면 다음과 같다. 먼저, 사용자 전체에 대한 평가데이터와 영화의 특징 벡터를 결합하고 정제 과정을 진행하여 테스트 셋과 검증 셋으로 분리한다. 테스트 셋은 그대로 추천 과정까지 진행하여 추천 결과 테이블을 도출하고 순위로 정렬된 테이블에 검증 셋을 사용하여 영화를 보았는지, 안 보았는지를 구분하는 참/거짓 테이블을 구축한다. 구축된 테이블에서 순위에 대해 퍼센티지 기준으로 병합한 테이블을 생성하여 모든 사용자의 X 축이 동일하도록 만들고, 이를 토대로 전체 사용자의 테이블을 합하여 하나의 퍼센티지 대 시청빈도 테이블을 구축한다.

메타 가중치를 적용하지 않은 채로 이 과정을 임의의 1000 명에 대하여 진행한 뒤 시각화 한 그래프는 아래 그림 3 과 같다. 그림은 추천의 순위가 내려갈수록 시청 빈도가 적어지는 우하향 곡선 형태의 그래프가 도출되며 메타 가중치를 적용되지 않더라도 추천 과정이 유의미하게 적용 된 것을 볼 수 있다.

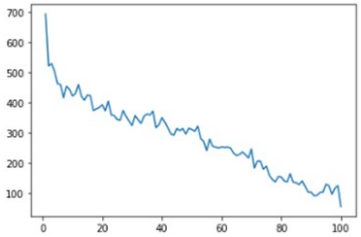

그림 3. 임의 1000명의 사용자를 대상으로 한 검증 결과
(x: 순위 퍼센티지, y: 시청빈도)

메타 가중치의 학습 과정은 상기한 검증 과정에 완벽한 우하향 직선의 패턴을 정답으로 설정한 손실함수와, 기계학습 기술의 일종인 경사 하강법(Gradient descent)을 활용하며 그림 4 와 같이 동작한다.

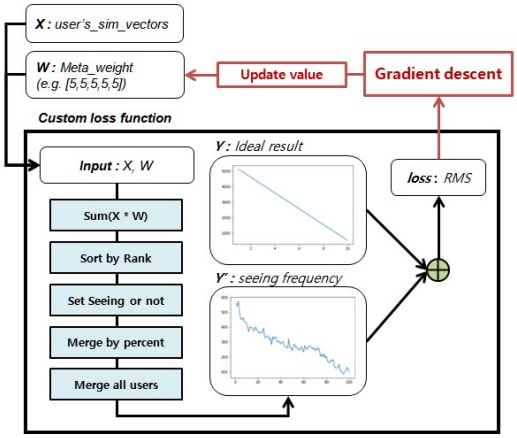

그림 4. 경사 하강법을 활용한 메타 가중치 학습 과정

1 부터 100 까지 퍼센티지화와 200 번의 반복을 통해 학습 한 결과 loss 값이 약 32%가량, 평균편차는 약 20% 가량 개선되었음을 확인하였다. 또한, 1000 명의 임의 샘플링 과정에서 오차가 발생하는 부분을 감안하여 20 번의 임의 샘플링을 진행해 본 결과 약 31.8±3.2%로 측정되었다.

이러한 학습 과정과 메타 가중치는 용도에 따라 다른 의도로 사용 될 수 있는데 특일이나 기상에 대한 적용이 그 예이다. 제안하는 추천 시스템에서 일자가 포함된 사용자 전체 평가 데이터에 특일이나 기상 등을 결합 할 수 있는데, 결합 된 외부 요인에 대하여 개인이 아닌 집단의 단위로 선호 벡터를 만들고 메타 가중치에 대한 학습을 진행한다면 결합 된 외부요인과 어울리는 영화를 추천할 수 있을 것이다[4]. 즉 크리스마스에 어울리는 영화나 비 오는 날 어울리는 영화 등의 추천이 가능해지는 것이다.

3. 결론

본 논문은 메타 가중치라는 요소를 활용한 개인화 추천시스템을 제안하였다. 본 연구의 제안하는 검증 모델을 통해 실험한 결과 선정된 5 개의 메타 데이터가 실제 영화의 특징을 잘 반영한다는 점을 확인 할 수 있었다.

최종적으로 도출된 메타 가중치는 학습마다 편차가 존재했으며, 대부분의 결과에서 장르가 낮은 가중치로 학습되는 경향을 보였다. 이는 여러 방향으로 해석 할 수 있다. 장르가 실제로 사용자의 선호 기준에 있어 줄거리나 감독, 배우, 인지도보다 영향을 적게 미친다거나, 사용자들이 가진 기록의 형태가 이미 선호하는 장르의 영화가 많은 상태로 저장되어 있다던가 하는 것이 그 예이다.

또한, 검증의 결과로 우하향 패턴의 곡선이 나오면 추천이 잘

되었다고 할 수 있지만 완벽한 우하향의 직선이 과연 이상적인 패턴을 의미할 수 있는지에 대해서는 다소 고민이 필요해 보인다.

참고문헌

[1] Jieun Son, et al. "Review and Analysis of Recommender Systems." Journal of the Korean Institute of Industrial Engineers 41.2 (2015): 185-208.

[2] Harper, F. Maxwell, and Joseph A. Konstan. "The movielens datasets: History and context." Acm transactions on interactive intelligent systems (tiis) 5.4 (2015): 1-19.

[3] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[4] An, Hyeon-woo, and Nammee Moon. "Design of recommendation system for tourist spot using sentiment analysis based on CNN-LSTM." Journal of Ambient Intelligence and Humanized Computing (2019): 1-11.