Matrix Factorization Techniques for Recommender Systems (2009)

태그 RecSys

논문 리뷰: Matrix Factorization Techniques for Recommender Systems

저자: Yehuda Koren (Yahoo Research), Robert Bell & Chris Volinsky (AT&T Labs-Research)

출처: IEEE Computer, 2009년 8월

---

1. 서론

오늘날 온라인 쇼핑몰과 콘텐츠 플랫폼은 과거에는 상상할 수 없었던 규모의 상품과 서비스를 제공하며, 개인의 다양한 취향을 만족시킬 기회를 열어주고 있습니다. 이러한 환경에서 사용자에게 가장 적합한 상품을 연결해주는 추천 시스템의 역할은 사용자 만족도와 충성도를 높이는 핵심 요소로 자리 잡았습니다.

특히 영화, 음악, TV 쇼와 같은 엔터테인먼트 상품 분야에서 추천 시스템은 더욱 빛을 발하며, 넷플릭스와 같은 기업들은 추천 시스템을 웹사이트의 핵심 기능으로 활용하고 있습니다.

이 논문은 넷플릭스 프라이즈 대회를 통해 그 우수성이 입증된 매트릭스 팩토리제이션 모델을 심도 있게 다룹니다. 이 기법은 기존의 최근접 이웃 방식보다 뛰어난 예측 정확도를 보여주며, 암묵적 피드백, 시간적 요소, 신뢰 수준 등 다양한 부가 정보를 통합할 수 있는 유연성을 제공합니다.

---

2. 추천 시스템의 주요 전략

추천 시스템은 크게 두 가지 전략을 기반으로 합니다.

2.1. 콘텐츠 필터링

각 사용자 또는 상품의 프로필을 생성하여 그 특성을 파악합니다. 예를 들어 영화 프로필은 장르, 배우, 흥행 성적 등의 속성을 포함할 수 있고, 사용자 프로필은 인구 통계학적 정보나 설문 응답을 포함할 수 있습니다.

이렇게 생성된 프로필을 바탕으로 사용자와 상품을 연결합니다. 다만, 외부 정보 수집이 필요하며, 수집이 어렵거나 불가능한 경우가 있다는 단점이 있습니다.

2.2. 협업 필터링

명시적인 프로필 생성 없이 과거 사용자 행동(예: 거래 내역, 상품 평점)만을 분석합니다. 사용자 간의 관계 및 상품 간의 상호 의존성을 분석하여 새로운 사용자-상품 연관성을 식별합니다. 도메인 지식이 필요 없어 범용적이며, 콘텐츠 필터링으로는 파악하기 어려운 미묘한 데이터 특성을 잡아낼 수 있다는 장점이 있습니다.

하지만 신규 사용자나 상품에 대한 추천이 어려운 콜드 스타트 문제가 발생할 수 있습니다.

---

3. 협업 필터링의 세부 기법

협업 필터링은 다시 최근접 이웃 방식과 잠재 요인 모델로 나뉩니다.

3.1. 최근접 이웃 방식

특정 상품에 대한 사용자 선호도를 해당 사용자가 평가한 '이웃' 상품들의 평점을 기반으로 평가하거나, 유사한 취향의 사용자들을 찾아 그들의 평가를 활용합니다.

3.2. 잠재 요인 모델

평점 패턴에서 추론된 20~100개 정도의 '요인'을 통해 사용자와 상품 모두의 특성을 파악하려 합니다. 이 요인들은 코미디 대 드라마, 액션의 양, 아동 지향성 등 명확한 차원일 수도 있고, 캐릭터 발전의 깊이나 독특함과 같이 덜 명확하거나 해석 불가능한 차원일 수도 있습니다.

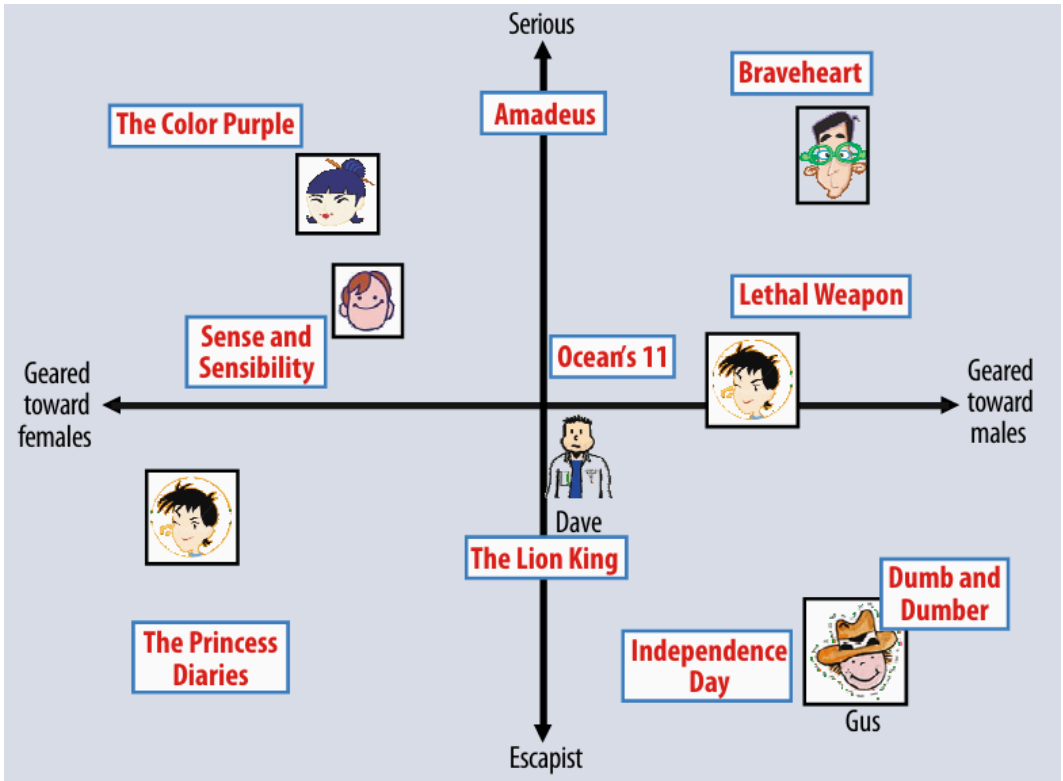

Figure 2. A simplified illustration of the latent factor approach, which characterizes both users and movies using two axes—male versus female and serious versus escapist.

4. Matrix Factorization 기법

잠재 요인 모델 중 가장 성공적인 구현 방법이 바로 Matrix Factorization입니다. 이 기법은 사용자와 상품을 평점 패턴으로부터 추론된 요인 벡터로 나타내며, 사용자와 상품 요인 간의 높은 일치도는 추천으로 이어집니다.

4.1. 기본 Matrix Factorization 모델

Matrix Factorization 모델은 사용자와 상품을 f 차원의 잠재 요인 공간에 매핑하여, 사용자-상품 간 상호작용을 해당 공간에서의 내적으로 모델링합니다.

상품 i는 벡터 q_i와 연관, 사용자 u는 벡터 p_u와 연관됩니다. q_i의 요소들은 상품이 해당 요인들을 얼마나 가지고 있는지를, p_u의 요소들은 사용자가 해당 요인들에 높은 값을 가지는 상품에 얼마나 관심 있는지를 나타냅니다. 이 두 벡터의 내적은 사용자 u와 상품 i간의 상호작용, 즉 상품 특성에 대한 사용자의 전반적인 관심을 포착하며, 이는 사용자 u의 상품 i에 대한 평점 r_{ui}의 추정치가 됩니다.

$$\hat{r}_{ui} = q_i^T p_u$$

가장 큰 과제는 각 상품과 사용자를 요인 벡터 q_i, p_u로 매핑하는 방법을 계산하는 것입니다. 이 매핑이 완료되면, 시스템은 위 공식을 사용하여 사용자가 어떤 상품에 대해 어떤 평점을 줄지 쉽게 추정할 수 있습니다. 이 모델은 정보 검색 분야에서 잠재 의미 요인을 식별하는 데 사용되는 특이값 분해(SVD)와 밀접하게 관련되어 있습니다.

그러나 사용자-상품 평점 행렬의 높은 희소성(결측값 비율이 높음) 때문에 전통적인 SVD 적용은 어렵습니다. 초기 시스템들은 결측값을 채워 넣어 밀집 행렬을 만드는 방식을 사용했지만, 이는 비용이 많이 들고 부정확한 값 주입은 데이터를 왜곡시킬 수 있습니다. 따라서 최근 연구들은 관찰된 평점만을 직접 모델링하면서 정규화된 모델을 통해 과적합을 방지하는 방식을 제안합니다.

요인 벡터 p_u와 q_i를 학습하기 위해, 시스템은 알려진 평점 세트에 대한 정규화된 제곱 오차를 최소화합니다.

$$\min_{q_*, p_*} \sum_{(u,i) \in \kappa} (r_{ui} - q_i^T p_u)^2 + \lambda (\|q_i\|^2 + \|p_u\|^2)$$

여기서 κ는 r_{ui}가 알려진 (u, i) 쌍의 집합(훈련 세트)입니다. λ는 정규화의 정도를 제어하는 상수로, 교차 검증을 통해 결정됩니다. 정규화는 학습된 파라미터의 크기에 페널티를 부과하여 과적합을 피하고, 관찰된 평점을 일반화하여 미래의 알려지지 않은 평점을 예측하는 것을 목표로 합니다.

5. 학습 알고리즘

위의 오차 함수를 최소화하기 위한 두 가지 주요 접근 방식은 다음과 같습니다.

5.1. 확률적 경사 하강법 (SGD)

훈련 세트의 모든 평점을 반복적으로 확인합니다. 각 평점에 대해 예측값 $\hat{r}_{ui}$을 계산하고 예측 오차를 구합니다. 그런 다음, 오차 기울기의 반대 방향으로 학습률 γ에 비례하는 크기만큼 파라미터 q_i와 p_u를 수정합니다.

$$q_i \leftarrow q_i + \gamma \cdot (e_{ui} \cdot p_u - \lambda \cdot q_i)$$

$$p_u \leftarrow p_u + \gamma \cdot (e_{ui} \cdot q_i - \lambda \cdot p_u)$$

구현이 쉽고 실행 시간이 비교적 빠르다는 장점이 있습니다.

5.2. 교대 최소 제곱법 (ALS)

q_i와 p_u가 모두 미지수이므로 오차 함수는 볼록하지 않습니다. 그러나 하나를 고정하면 최적화 문제는 이차식이 되어 최적으로 풀 수 있습니다. ALS는 q_i를 고정하고 p_u를 최적화한 다음, p_u를 고정하고 q_i를 최적화하는 과정을 반복합니다.

각 단계에서 오차를 감소시키며 수렴합니다. 일반적으로 SGD가 ALS보다 쉽고 빠르지만, ALS는 병렬화가 가능하거나 암묵적 데이터 중심 시스템에서 유리합니다. 암묵적 데이터의 경우 훈련 세트가 희소하지 않아 SGD처럼 각 훈련 사례를 반복하는 것이 비실제적일 수 있는데, ALS는 이러한 경우를 효율적으로 처리할 수 있습니다.

---

6. 모델 확장 및 개선

기본적인 매트릭스 팩토리제이션 모델은 다양한 데이터 측면과 특정 애플리케이션 요구 사항을 처리하기 위해 유연하게 확장될 수 있습니다.

6.1. 편향 추가

관찰된 평점 변화의 상당 부분은 사용자와 상품 간의 상호작용보다는 사용자 또는 상품 자체와 관련된 효과, 즉 편향 또는 절편 때문입니다. 예를 들어, 어떤 사용자는 다른 사용자보다 평점을 후하게 주는 경향이 있고, 어떤 상품은 다른 상품보다 높은 평점을 받는 경향이 있습니다. 이러한 편향을 모델에 명시적으로 포함하면 예측 정확도를 높일 수 있습니다.

평점 r_{ui}에 대한 편향은 전체 평균 평점 mu, 상품 i의 편향 b_i, 사용자 u의 편향 b_u의 합으로 근사할 수 있습니다. $b_{ui} = \mu + b_i + b_u$. 이를 반영한 예측 평점은 다음과 같습니다.

$\hat{r}_{ui} = \mu + b_i + b_u + q_i^T p_u$. 새로운 오차 함수는 다음과 같이 정의됩니다.

$\min_{p_*, q_*, b_*} \sum_{(u,i) \in \kappa} (r_{ui} - \mu - b_u - b_i - p_u^T q_i)^2 + \lambda (\|p_u\|^2 + \|q_i\|^2 + b_u^2 + b_i^2)$

6.2. 추가 입력 소스

콜드 스타트 문제를 완화하기 위해 추가적인 사용자 정보를 통합할 수 있습니다. 대표적인 것이 암묵적 피드백입니다.사용자가 명시적인 평점을 제공하지 않더라도 구매 내역, 검색 패턴, 브라우징 기록 등의 행동 정보를 수집하여 선호도를 파악할 수 있습니다.

예를 들어, 사용자 u가 암묵적으로 선호도를 표현한 상품 집합을 $N(u)$라고 할 때, 새로운 상품 요인 벡터 x_i를 도입하여 사용자를 $\sum_{i \in N(u)} x_i$ (또는 정규화된 형태 $|N(u)|^{-0.5} \sum_{i \in N(u)} x_i$)로 특징지을 수 있습니다.

또한, 인구 통계학적 정보와 같은 사용자 속성도 활용될 수 있습니다. 사용자 u와 관련된 속성 집합을 $A(u)$라고 할 때, 각 속성에 해당하는 요인 벡터 y_a를 사용하여 사용자를 $\sum_{a \in A(u)} y_a$로 나타낼 수 있습니다.

이러한 모든 정보 소스를 통합한 향상된 사용자 표현을 사용하는 예측 모델은 다음과 같습니다.

$\hat{r}_{ui} = \mu + b_i + b_u + q_i^T \left(p_u + |N(u)|^{-0.5} \sum_{j \in N(u)} x_j + \sum_{a \in A(u)} y_a \right)$

6.3. 시간적 동역학

상품의 인기도나 사용자의 취향은 시간에 따라 변합니다. Matrix Factorization은 이러한 시간적 효과를 모델링하는 데 적합합니다. 구체적으로 상품 편향 b_i(t), 사용자 편향 $b_u(t)$, 사용자 선호도 $p_u(t)$가 시간에 따라 변하는 것을 고려할 수 있습니다.

반면, 상품 자체의 특성 q_i는 정적이라고 가정합니다. 이를 반영한 동적 예측 규칙은 다음과 같습니다.

$\hat{r}_{ui}(t) = \mu + b_i(t) + b_u(t) + q_i^T p_u(t)$

6.4. 다양한 신뢰 수준을 가진 입력

모든 관찰된 평점이 동일한 가중치나 신뢰도를 갖는 것은 아닙니다. 예를 들어, 대규모 광고는 특정 상품에 대한 투표에 영향을 미칠 수 있으며, 이는 장기적인 특성을 제대로 반영하지 못할 수 있습니다.

또한, 암묵적 피드백 기반 시스템에서는 사용자의 정확한 선호도 수준을 정량화하기 어렵습니다. 대신 "아마도 좋아할 것이다" 또는 "아마도 관심 없을 것이다"와 같은 이진 표현을 사용하며, 이때 각 관찰에 대한 신뢰도 점수를 부여하는 것이 유용합니다.

신뢰도는 행동 빈도(예: 특정 쇼 시청 시간, 특정 상품 구매 빈도)로부터 얻을 수 있습니다. 관찰된 평점 r_{ui}에 대한 신뢰도를 c_{ui}라고 할 때, 수정된 오차 함수는 다음과 같습니다.

$\min_{p_*, q_*, b_*} \sum_{(u,i) \in \kappa} c_{ui} (r_{ui} - \mu - b_u - b_i - p_u^T q_i)^2 + \lambda (\|p_u\|^2 + \|q_i\|^2 + b_u^2 + b_i^2)$

---

7. 넷플릭스 프라이즈 대회와 매트릭스 팩토리제이션

2006년 넷플릭스는 자사 추천 시스템의 성능 개선을 위한 대회를 개최했습니다. 약 50만 명의 익명 고객과 17,000편 이상의 영화에 대한 1억 건이 넘는 평점 데이터가 공개되었습니다.

이 논문의 저자들이 속한 팀(BellKor)은 여기에 설명된 방법들의 변형을 사용한 100개 이상의 예측기 세트로 2007년과 2008년 Progress Prize를 수상했으며, 결국 100만 달러의 최종 상금을 획득하는 데 핵심적인 역할을 했습니다. 넷플릭스 사용자-영화 행렬을 분해함으로써 영화 선호도를 예측하는 데 가장 설명력이 높은 차원들을 발견할 수 있었습니다.

예를 들어, 첫 번째 요인 벡터는 한쪽에는 남성 또는 청소년 관객을 대상으로 하는 저급 코미디와 공포 영화가, 다른 쪽에는 진지한 분위기를 풍기거나 여성 주연이 강한 드라마나 코미디가 위치했습니다. 두 번째 요인 벡터는 위쪽에는 독립적이고 비평가들의 호평을 받은 독특한 영화가, 아래쪽에는 주류의 정형화된 영화가 위치했습니다. 이러한 잠재 요인 분석은 영화들의 미묘한 특성과 사용자들의 다양한 취향을 포착하는 데 유용함을 보여주었습니다.

논문에서 제시된 다양한 모델과 파라미터 수에 따른 RMSE 변화를 보면, 모델의 복잡성(더 많은 파라미터 사용, 즉 요인 모델의 차원 증가)이 증가할수록 정확도가 향상되는 것을 확인할 수 있습니다.

특히 시간적 요소를 모델링하는 것이 데이터의 중요한 시간적 효과를 포착하는 데 매우 중요했습니다.

---

8. 결론

매트릭스 팩토리제이션 기법은 협업 필터링 추천 시스템 내에서 지배적인 방법론으로 자리 잡았습니다.

넷플릭스 프라이즈와 같은 대규모 데이터셋을 통한 경험은 이 기법이 고전적인 최근접 이웃 방식보다 우수한 정확도를 제공함을 보여주었습니다. 동시에, 상대적으로 쉽게 학습할 수 있는 작고 메모리 효율적인 모델을 제공합니다. 더욱이, 이러한 기법들은 다중 피드백 형태, 시간적 동역학, 신뢰 수준과 같은 데이터의 여러 중요한 측면을 자연스럽게 통합할 수 있다는 점에서 매우 편리합니다.

이 논문은 매트릭스 팩토리제이션의 기초부터 다양한 확장 방법론까지 체계적으로 설명하며 추천 시스템 연구 및 개발에 중요한 지침을 제공합니다.