**📚 추천 시스템 비교 및 분석 종합 요약**

**I. 기본 모델 (Base Models)**

| **모델명** | **사용 이유 및 목적 (Why)** | **작동 방법 (How)** | **주요 장단점 및 성능 평가** |
| --- | --- | --- | --- |
| **CF 단독 모델 (협업 필터링)** | **패턴 학습의 기본:** 17만 명의 유저들이 50개 게임에 남긴 **잠재적인 상호작용 패턴**을 학습하여 추천합니다. | **Matrix Factorization (MF):** 유저와 아이템을 잠재 요인 벡터($P, Q$)로 분해한 후, 이 벡터의 내적($P \cdot Q$)으로 예측 평점을 산출합니다. | **성능 저조:** 50개 아이템 환경에서는 데이터가 너무 희소하여 **취향 패턴 학습에 실패**하고, ID: 100 같은 노이즈나 일반적인 인기에 휘둘리는 불안정한 결과를 낳았습니다. |
| **Hybrid Filtering** | **안정성 확보:** CF의 불안정한 예측을 **CBF의 확실한 콘텐츠 연관성**으로 보강하기 위해 사용합니다. | **가중 평균 (Mixing):** CF 점수와 CBF 점수를 따로 계산한 뒤, **CBF에 70%의 높은 가중치**를 부여하여 최종 추천 점수를 산출합니다. | **성능 우수:** **가장 합리적인 솔루션**으로 입증되었습니다. CBF가 유저의 명확한 취향(예: Half-Life 시리즈)을 확실하게 저격하여, 제한된 데이터 환경에서 높은 안정성과 품질을 보장합니다. |
| **콘텐츠 임베딩** | **CBF 지능 강화:** CBF의 유사도 계산 기준을 단순 키워드(TF-IDF)에서 \*\*더 깊은 콘텐츠 의미(Title, Developer, Publisher)\*\*로 확장하여 유사도 측정의 정확도를 높입니다. | **(시뮬레이션):** title, developer, publisher 정보를 통합하여 TF-IDF 벡터를 생성하고, 이를 기반으로 코사인 유사도를 측정합니다. (실제 임베딩은 Word2Vec/BERT 사용) | **장점:** 아이템 자체 속성을 심층적으로 파악하여, **데이터 희소성의 영향을 받지 않는** CBF 컴포넌트의 성능을 비약적으로 개선합니다. |

**II. 고급 머신러닝 기법 (Advanced ML Techniques)**

| **기법명** | **사용 이유 및 목적 (Why)** | **작동 방법 (How)** | **주요 장단점 및 성능 평가** |
| --- | --- | --- | --- |
| **1. CF with Biases** | **CF의 기반 다지기:** CF 예측의 정확도를 해치는 \*\*'아이템의 일반적인 인기'\*\*나 \*\*'유저의 평균적인 평점 습관'\*\*이라는 편향을 분리하여 CF 모델의 안정성을 높이기 위함입니다. | **편향 항 추가:** 기존 MF 공식에 **전역 평균($\mu$), 유저 편향($b\_u$), 아이템 편향($b\_i$)** 항을 추가하여 예측합니다. 이 편향 항들을 잠재 요인($P, Q$)과 함께 학습합니다. | **장점:** 구현이 쉽고, MF 예측을 **더 이성적으로** 만듭니다. 노이즈의 영향을 줄여 순수한 취향 패턴 학습에 집중하도록 돕습니다. |
| **2. Factorization Machines (FM)** | **궁극의 피처 융합:** CF와 CBF의 정보를 따로 계산하지 않고, **하나의 모델로 통합하여 모든 피처 쌍의 상호작용**을 학습합니다. | **통합 피처셋:** 유저/아이템 ID (CF)와 게임 제목/속성 (CBF)을 모두 **희소 원-핫 인코딩** 및 **TF-IDF 벡터**로 변환한 뒤, 이 거대한 통합 행렬을 모델의 입력으로 사용합니다. | **장점:** **이론상 최고 성능**을 자랑합니다. 유저의 취향과 아이템의 콘텐츠 속성 간의 복잡한 규칙(예: '이 유저는 Valve가 만든 게임에만 평점을 잘 준다')을 학습하여 희소성 문제를 가장 정교하게 극복합니다. |

**💡 최종 결론: 제한된 데이터 환경에서의 전략**

제한된 **50개 아이템** 데이터셋 환경에서는 **CF의 패턴 학습 능력은 크게 약화**됩니다.

따라서 가장 현명한 전략은 다음과 같습니다.

1. **안정성 확보:** **CBF** (콘텐츠 유사성)에 높은 가중치를 부여하는 **하이브리드 모델**을 사용해야 합니다.
2. **기반 강화:** **CF with Biases**와 \*\*콘텐츠 임베딩(CBF 강화)\*\*을 사용하여 두 기반 모델의 예측 품질을 최대한 끌어올립니다.
3. **최종 목표:** 데이터 규모가 커지거나 리소스가 충분해지면, 모든 정보를 완벽하게 통합하는 **Factorization Machines**로 전환하여 시스템을 고도화합니다.

**좀 더 자세히**

**1. Collaborative Filtering (CF) - 패턴 분석 및 취향 발견**

| **상세 역할** | **작동 원리** | **50개 아이템 환경에서의 의미** |
| --- | --- | --- |
| **패턴 분석가 (Pattern Analyst)** | **Matrix Factorization (행렬 분해):** 유저-아이템 평점 행렬($R$)을 유저 잠재 요인 행렬($P$)과 아이템 잠재 요인 행렬($Q$)로 분해합니다. 유저 $u$의 벡터($P\_u$)와 아이템 $i$의 벡터($Q\_i$)가 비슷할수록 높은 예측 평점($\hat{r}\_{ui}$)을 내놓습니다. | **잠재력 vs. 현실:** 아이템이 50개뿐이라 학습할 패턴 자체가 극도로 희소합니다. $P$와 $Q$가 유저의 진짜 취향 대신 데이터의 노이즈나 인기도만 학습할 위험이 큽니다. |
| **핵심 임무** | \*\*"유사한 취향을 가진 다른 유저들이 당신이 아직 경험하지 않은 아이템 중 무엇을 선호하는가?"\*\*를 예측합니다. |  |

**2. CBF (Content-Based Filtering) / 콘텐츠 임베딩 - 안정성 확보 및 개인 취향 관리**

| **상세 역할** | **작동 원리** | **50개 아이템 환경에서의 의미** |
| --- | --- | --- |
| **개인 취향 관리자 (Personal Taste Keeper)** | **TF-IDF 및 코사인 유사도:** 아이템의 콘텐츠(제목, 장르 등)를 벡터화(특징 추출)한 후, 유저가 좋아했던 아이템의 벡터와 다른 모든 아이템 벡터 간의 **유사도**를 측정하여 추천 점수를 산출합니다. | **안전 닻 역할:** CF와 달리 유저 수가 아닌 **아이템 자체 속성**에 의존하므로, 데이터가 아무리 희소해도 예측의 안정성이 보장됩니다. 하이브리드 모델의 **안전 닻** 역할을 수행합니다. |
| **콘텐츠 특징 강화** | **콘텐츠 임베딩 (Word2Vec, BERT):** 제목, 리뷰 등의 텍스트를 단순 키워드가 아닌 **의미론적 벡터**로 변환하여 CBF의 유사도 계산 정확도를 획기적으로 향상시킵니다. |  |

**3. Hybrid Filtering (가중 평균) - 리스크 관리 및 최종 의사 결정**

| **상세 역할** | **작동 원리** | **50개 아이템 환경에서의 의미** |
| --- | --- | --- |
| **리스크 관리자 (Risk Manager)** | **가중 평균 (Weighted Sum):** CF의 예측 점수($S\_{CF}$)와 CBF의 유사도 점수($S\_{CBF}$)를 단순 합산하지 않고, 시스템의 안정성을 보장하는 쪽에 더 큰 가중치를 부여합니다. $\text{Score} = (0.3 \cdot S\_{CF}) + (0.7 \cdot S\_{CBF})$ | **CBF 편향 의도:** CF의 불안정성이 매우 높기 때문에, CBF에 높은 가중치(0.7)를 부여함으로써 \*\*"CF의 예측을 참고하되, CBF의 안정성을 절대 포기하지 않겠다"\*\*는 전략적 의사결정을 수행합니다. |
| **핵심 임무** | 두 모델의 장점을 활용하여, **가장 신뢰할 수 있는 최종 추천 순위**를 확정합니다. |  |

**4. CF with Biases (편향 추가된 CF) - 모델 예측의 정제**

| **상세 역할** | **작동 원리** | **50개 아이템 환경에서의 의미** |
| --- | --- | --- |
| **노이즈 필터 (Noise Filter)** | **편향 항 추가:** 평점($\hat{r}\_{ui}$)을 예측할 때, 유저의 취향($P\_u Q\_i^T$) 외에 전역 평균($\mu$), 아이템 인기도($b\_i$), 유저 평점 습관($b\_u$)을 따로 분리하여 모델링합니다. $\hat{r}\_{ui} = \mu + b\_u + b\_i + P\_u Q\_i^T$ | **순수 취향 고립:** $\mu, b\_u, b\_i$가 아이템의 일반적인 인기를 대신 설명해주므로, 나머지 잠재 요인($P, Q$)은 **인기가 제거된 순수한 유저의 개인적인 취향**만을 학습하게 되어 모델의 해석력과 안정성이 높아집니다. |
| **핵심 임무** | \*\*"이 아이템의 평점이 높은 것이 정말 내 취향 때문인가, 아니면 그냥 인기 게임이기 때문인가?"\*\*를 구분하여 CF의 학습 효율을 높입니다. |  |

**5. Factorization Machines (FM) - 피처 통합 및 컨텍스트화**

| **상세 역할** | **작동 원리** | **50개 아이템 환경에서의 의미** |
| --- | --- | --- |
| **최상위 통합 예측기 (Ultimate Feature Fusion)** | **피처 간 상호작용 학습:** CF 피처(유저/아이템 ID)와 CBF 피처(콘텐츠 속성)를 모두 입력받아, 피처 하나하나의 영향력뿐만 아니라 **모든 피처 쌍 간의 상호작용($\text{ID} \times \text{Keyword}$)을 명시적으로 학습**합니다. | **데이터 구멍 메우기:** 50개 아이템 데이터에서 CF만으로는 유저의 잠재 요인($P$)을 알기 어렵습니다. FM은 \*\*"이 유저는 'Valve'가 만든 게임($\text{CBF}$ 피처)을 좋아한다"\*\*는 규칙을 학습함으로써, 부족한 $\text{CF}$ 정보를 $\text{CBF}$ 피처로 **대신 설명**하며 예측의 정확도를 최고로 끌어올립니다. |
| **핵심 임무** | \*\*"이 유저 ID의 이 콘텐츠 속성에 대한 반응"\*\*이라는 복잡한 문맥(Context)을 이해하여 최적의 예측을 생성합니다. |  |