9장. 추천 시스템

01 추천 시스템의 개요와 배경

추천 시스템의 유형

- 콘텐츠 기반 필터링
- 협업 필터링
  - 최근접 이웃 협업 필터링
  - 잠재 요인 협업 필터링

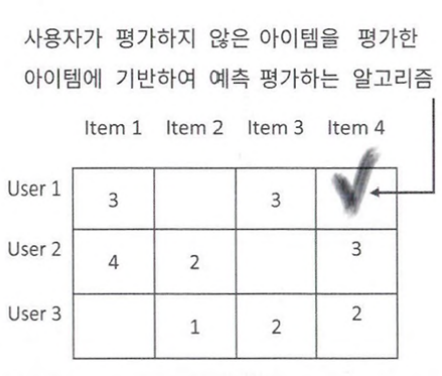
02 콘텐츠 기반 필터링 추천 시스템

- 사용자가 특정한 아이템을 매우 선호하는 경우, 그 아이템과 비슷한 콘텐츠를 가진 다른 아이템을 추천

03 최근접 이웃 협업 필터링

- 협업 필터링: 사용자 행동 양식만을 기반으로 추천을 수행

  *사용자 행동 양식: 사용자가 아이템에 매긴 평점 정보나 상품 구매 이력 등
- 사용자 행동 데이터를 기반으로 사용자가 아직 평가하지 않은 아이템을 예측 평가하는 것



- 최근접 이웃 필터링(메모리 협업 필터링)

- 사용자 기반: 유사한 다른 사용자를 Top-N으로 선정해 Top-N 사용자가 좋아하는 아이템 추천

	다크 나이트	인터스텔라	엣지오브 투모로우	프로메테우스	스타워즈 라스트제다이
사용자 A	5	4	4		
사용자 B	5	3	4	5	3
사용자 C	4	3	3	2	5

상호간 유사도 높음

사용자 A는 사용자 C보다 사용자 B와 영화 평점 측면에서 유사도가 높음. 따라서 사용자 A에게는 사용자 B가 재미있게 본 '프로메테우스'를 추천

- 아이템 기반: 사용자들의 아이템에 대한 평가 척도가 유사한 아이템을 추천

	사용자 A	사용자 B	사용자 C	사용자 D	사용자 E
다크 나이트	5	4	5	5	5
프로메테우스	5	4	4		5
스타워즈 라스트제다이	4	3	3		4

상호간 유사도 높음

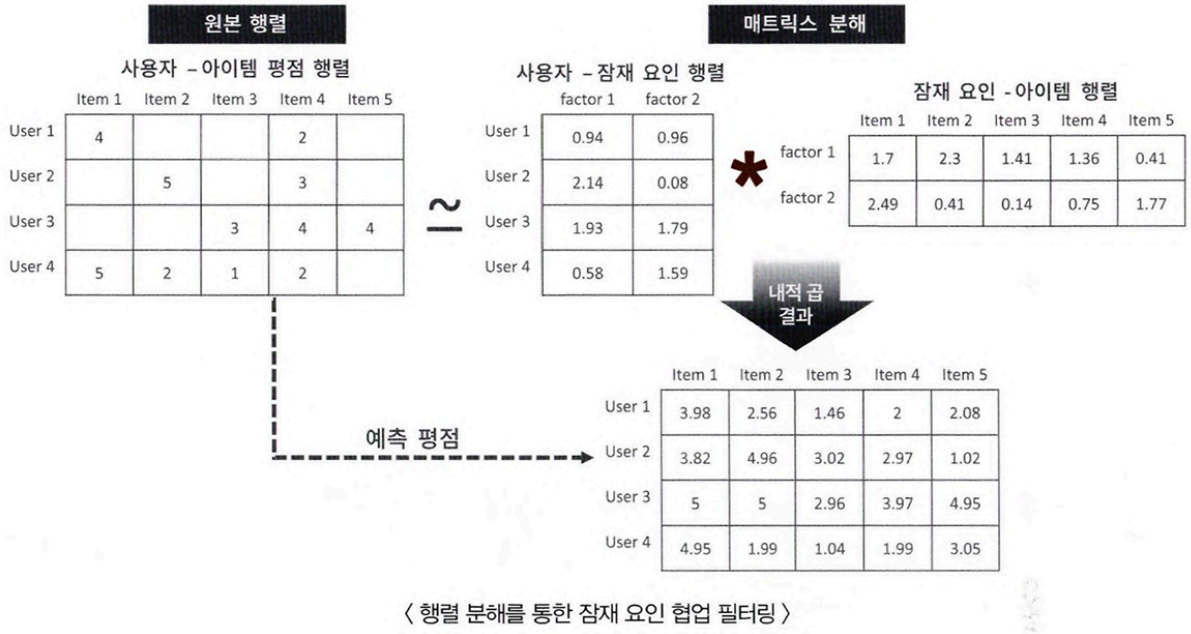
여러 사용자들의 평점을 기준으로 볼 때 '다크 나이트'와 가장 유사한 영화는 '프로메테우스'

비슷한 상품을 선호한다고 해서 취향이 비슷하다고 판단하기 어렵기 때문에, 일반적으로 아이템 기반 협업 필터링의 정확도가 더 높음

04 잠재 요인 협업 필터링

잠재 요인 협업 필터링의 이해

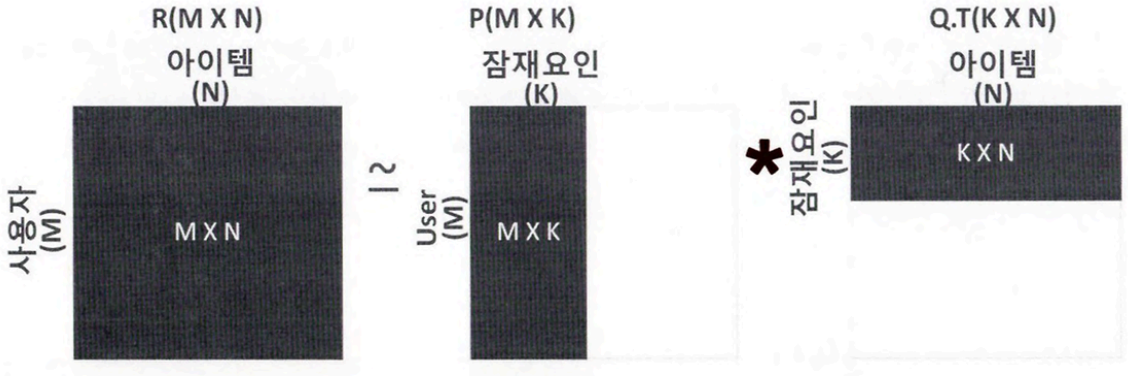
- 대규모 다차원 행렬을 차원 감소 기법으로 분해하는 과정에서 잠재 요인을 추출해, 추천 예측을 할 수 있는 기법



〈 행렬 분해를 통한 잠재 요인 협업 필터링 〉

행렬 분해의 이해

- 다차원의 매트릭스를 저차원 매트릭스로 분해하는 기법
- SVD, NMF 등이 있음



⇒ R = P*Q.T

M: 총 사용자 수

N: 총 아이템 수

K: 잠재 요인의 차원 수

R: M x N 차원의 사용자-아이템 평점 행렬

P: 사용자와 잠재 요인과의 관계 값을 가지는 M x K 차원의 사용자-잠재 요인 행렬

Q: 아이템과 잠재 요인과의 관계 값을 가지는 M x K 차원의 아이템-잠재 요인 행렬

Q.T: Q 매트릭스의 행과 열 값을 교환한 전치 행렬

확률적 경사 하강법을 이용한 행렬 분해

- P와 Q 행렬로 계산된 예측 R 행렬 값이 실제 R 행렬 값과 가장 최소의 오류를 가질 수 있도록 반복적인 비용 함수 최적화를 통해 P와 Q를 유추해내는 것

  전반적인 절차는 아래와 같다.

1. P와 Q를 임의의 값을 가진 행렬로 설정합니다.
2. P와 Q.T 값을 곱해 예측 R 행렬을 계산하고 예측 R 행렬과 실제 R 행렬에 해당하는 오류 값을 계산합니다.
3. 이 오류 값을 최소화할 수 있도록 P와 Q 행렬을 적절한 값으로 각각 업데이트합니다.
4. 만족할 만한 오류 값을 가질 때까지 2, 3번 작업을 반복하면서 P와 Q 값을 업데이트해 근사화합니다.

08 파이썬 추천 시스템 패키지 - Surprise

Surprise 패키지 소개

- 파이썬 기반의 추천 시스템 구축을 위한 전용 패키지(사이킷런은 제공하지 않음)
- 다양한 추천 알고리즘을 쉽게 적용 가능하며, 사이킷런의 핵심 API와 유사한 API명으로 작성되어 사용하기 편리하다는 장점을 가지고 있음

Surprise 주요 모듈 소개

API	내용
Dataset.load_builtin	FTP 서버에서 데이터를 내려받음
Dataset.load_from_file	OS 파일에서 데이터를 로딩
Dataset.olad_from_df	판다사의 DataFrame에서 데이터 로딩

- Dataset → OS 파일 데이터를 Surprise 데이터 세트로 로딩 → 판다스 DataFrame에서 Surprise 데이터 세트로 로딩

Surprise 추천 알고리즘 클래스

- Surprise에서 추천 예측을 위해 자주 사용되는 추천 알고리즘 클래스

클래스명	설명
SVD	행렬 분해를 통한 잠재 요인 협업 필터링을 위한 SVD 알고리즘.
KNNBasic	최근접 이웃 협업 필터링을 위한 KNN 알고리즘.
BaselineOnly	사용자 Bias와 아이템 Bias를 감안한 SGD 베이스라인 알고리즘.

- SVD 클래스의 입력 파라미터

파라미터명	내용
n_factors	잠재 요인 K의 개수. 디폴트는 100. 커질수록 정확도가 높아질 수 있으나 과적합 문제가 발생할 수 있습니다.
n_epochs	SGD(Stochastic Gradient Descent) 수행 시 반복 횟수. 디폴트는 20.
biased (bool)	베이스라인 사용자 편향 적용 여부이며, 디폴트는 True입니다.

베이스라인 평점

- 베이스라인: 개인의 성향을 반영해 아이템 평가에 편향성 요소를 반영하여 평점을 부과하는 것
- 보통 (전체 평균 평점 + 사용자 편향 점수 + 아이템 편향 점수)로 계산됨
  - 전체 평균 평점 = 모든 사용자의 아이템에 대한 평점을 평균한 값
  - 사용자 편향 점수 = 사용자별 아이템 평점 평균 값 - 전체 평균 평점
  - 아이템 편향 점수 = 아이템별 평점 평균 값 - 전체 평균 평점

교차 검증과 하이퍼 파라미터 튜닝

- 교차 검증을 위해 cross_validate(), 하이퍼 파라미터 튜닝을 위해 GridSearchCV 클래스를 이용