SPHERE

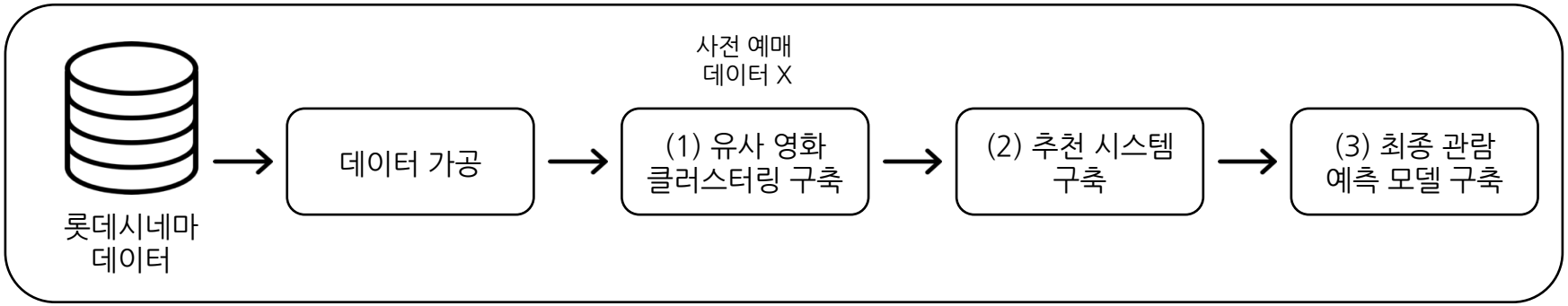
신규 영화 추천 시스템 알고리즘 정의서

tan.D

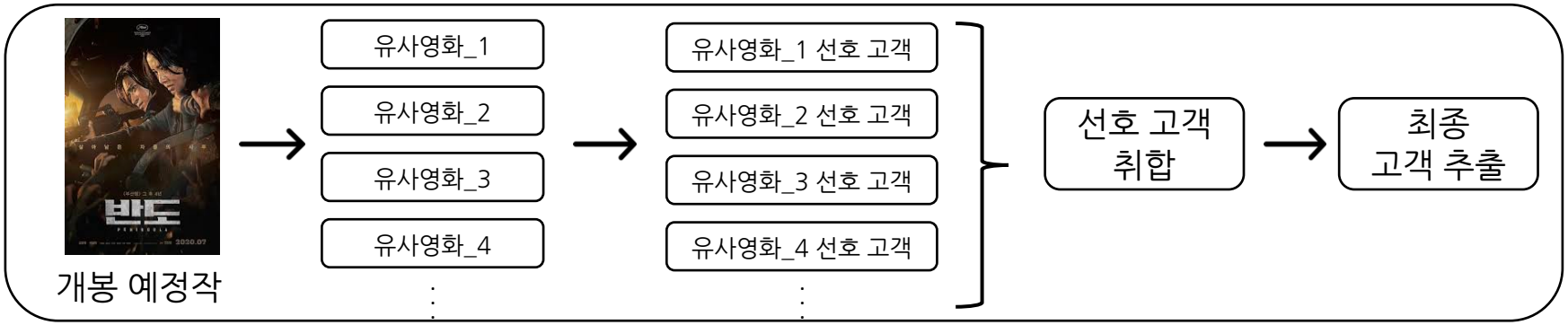
1. 프로젝트 과정

- 총 3가지 주요 과정
  - 유사 영화 클러스터링
  - 추천 시스템 : 유사영화별 선호 고객 추출 및 취합
  - 최종 관람 예측 모델 : 선호를 기반으로 추출된 고객 중 최근 관람 이력, 고객 정보등을 바탕으로 최종 관람 고객 추출

■ 프로젝트 과정



■ 실제 프로젝트 테스팅 과정



- 1) 클러스터링

(1) 개요
: 주어진 데이터들의 특성을 고려해 데이터 집단(클러스터)을 정의하고 데이터 집단을 대표할 수 있는 대표점을 찾는 것

(2) 목표
: 과거 개봉작들의 특성(영화 정보 - 장르, 줄거리, 배급사 등)을 고려해 여러 개의 집단으로 그룹화한 후, 관람하지 않은 신규 영화와 유사한 그룹(영화 목록)을 예측

< 예시 >

영화 명	장르	예상 흥행	배급사	그룹 명
A	액션	상	E사	그룹 1
B	범죄	상	E사	
C	호러	중	G사	그룹 2
D	호러	중	L사	
삼진그룹 영어토익반	드라마	중	L사	?

Q. 영화 '삼진그룹 영어토익반'과 유사한 영화 그룹은?

2. 유사 영화 추천

SPHERE

- 1) 클러스터링

1. 영화별 특징들을 숫자로 변환한 하나의 행렬 생성
   - 숫자로 변환하는 기법은 인코딩을 사용

2. 전체 개봉 영화들 중에서 무작위로 N개의 대표 영화를 추출
   - 추출된 N개의 대표 영화들이 각 그룹(클러스터)들의 대표로 설정

3. N개의 대표 영화를 제외한 모든 영화들을 각 그룹들의 대표들과 거리(유사도)를 계산
   - 각 그룹들 중에서 가장 높은 유사도를 받은 곳으로 배치
   - 유사도는 데이터의 특성들 중에서 일치하는 개수로 표현

4. 앞선 1~3단계를 반복해서 최적의 그룹 결과 생성

- 1) 클러스터링

▪ Step ① 영화별 특징들을 숫자로 변환한 하나의 행렬 생성

: 주어진 영화 정보 데이터와 수집 데이터 활용해,
유사한 영화들 간에 그룹을 형성하고자 합니다.

<변환 전 데이터>

	장르	예상 흥행	배급사
A	멜로	상	L사
B	드라마	중	A사
C	코미디	상	A사
D	SF	중	B사
삼진그룹 영어토익반	드라마	중	B사

영화

수치화 →

① <변환 후 데이터>

	장르	예상 흥행	배급사
A	0	0	2
B	1	1	0
C	2	0	0
D	4	1	1
삼진그룹 영어토익반	1	1	1

① <변환 후 데이터>

: 위 예시와 같이 장르, 예상 흥행, 배급사 등의 문자 형태를 숫자로 변환합니다. 다양한 변환 방법이 존재하지만, 가장 보편적으로 사용되는 인코딩 기법을 사용하였습니다. 인코딩 기법이란 대상마다 임의의 숫자를 부여하여 매칭시키는 방법입니다.
e.g. 영화의 장르에 멜로, 드라마, 코미디, 액션, SF 5개가 존재한다면, 다음과 같이 정의할 수 있습니다.

장르 : { 멜로 : 0, 드라마 : 1, 코미디 : 2, 액션 : 3, SF : 4 }
예상 흥행 : { 상 : 0, 중 : 1, 하 : 2 }
배급사 : { A사 : 0, B사 : 1, L사 : 2 }

* 실제로는 추가적인 특징들이 많으나, 편의상 장르, 흥행, 배급사로 한정하였습니다.

- 1) 클러스터링

▪ Step ② 전체 개봉 영화들 중에서 무작위로 N개의 대표 영화를 추출

: 추출된 N개의 대표 영화들이 각 그룹(클러스터)들의 대표로 설정합니다.

영화

	장르	예상 흥행	배급사
A	0	0	2
B	1	1	0
C	2	0	0
D	4	1	1
삼진그룹 영어토익반	1	1	1

대표 영화 추출 →

<임의의 그룹 생성>

영화 B 그룹 영화 D 그룹

- 1) 클러스터링

▪ Step ③ N개의 대표 영화를 제외한 모든 영화들을 각 그룹들의 대표들과 거리(유사도)를 계산

: 유사도 계산 후, 가장 유사한 그룹에 매칭을 시킵니다.

<입력 데이터>

영화

	장르	예상 흥행	배급사
A	0	0	2
B	1	1	0
C	2	0	0
D	4	1	1
삼진그룹 영어토익반	1	1	1

유사도 계산 →

영화 B 그룹 / 영화 D 그룹

	영화 B 그룹과의 유사도	영화 D 그룹과의 유사도
A	0	0
C	1	0
삼진그룹 영어토익반	2	2

<유사도 계산>

: 유사도를 계산하는 방법은 그룹의 대표 영화의 특징과 동일한 특징을 갖고 있는 개수를 사용하였습니다.
(장르가 동일한 지/예상 흥행도가 동일한 지/배급사가 동일한 지)

e.g. 유사도에 사용되는 특징의 개수는 장르, 예상 흥행, 배급사 3개로 한정됩니다.
위의 예시에서 A와 '삼진그룹 영어토익반'은 어느 그룹에도 속하지 않고, C는 영화 B 그룹에 속하게 됩니다.

- 다음과 같이 모든 영화가 명확하게 매칭이 안되는 경우, 대표 영화들을 다시 바꿔서 최적의 그룹을 찾습니다.

- 1) 클러스터링

▪ Step ④ 최적의 그룹 결과가 나올 때까지 Step ① ~ ③ 을 반복 실행

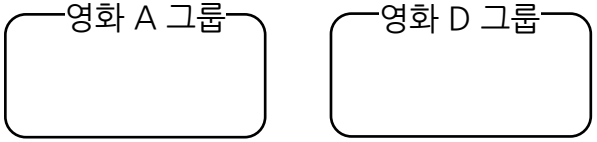
: 최적의 그룹 결과란 모든 영화들이 그룹들 중 한 곳에 모두 매칭이 된 것을 의미합니다.

〈입력 데이터〉

영화

	장르	예상 흥행	배급사
A	0	0	2
B	1	1	0
C	2	0	0
D	4	1	1
삼진그룹 영어토익반	1	1	1

유사도 계산 →



	영화 A 그룹과의 유사도	영화 D 그룹과의 유사도
B	0	1
C	1	0
삼진그룹 영어토익반	0	2

〈최적의 그룹 결과 예시〉

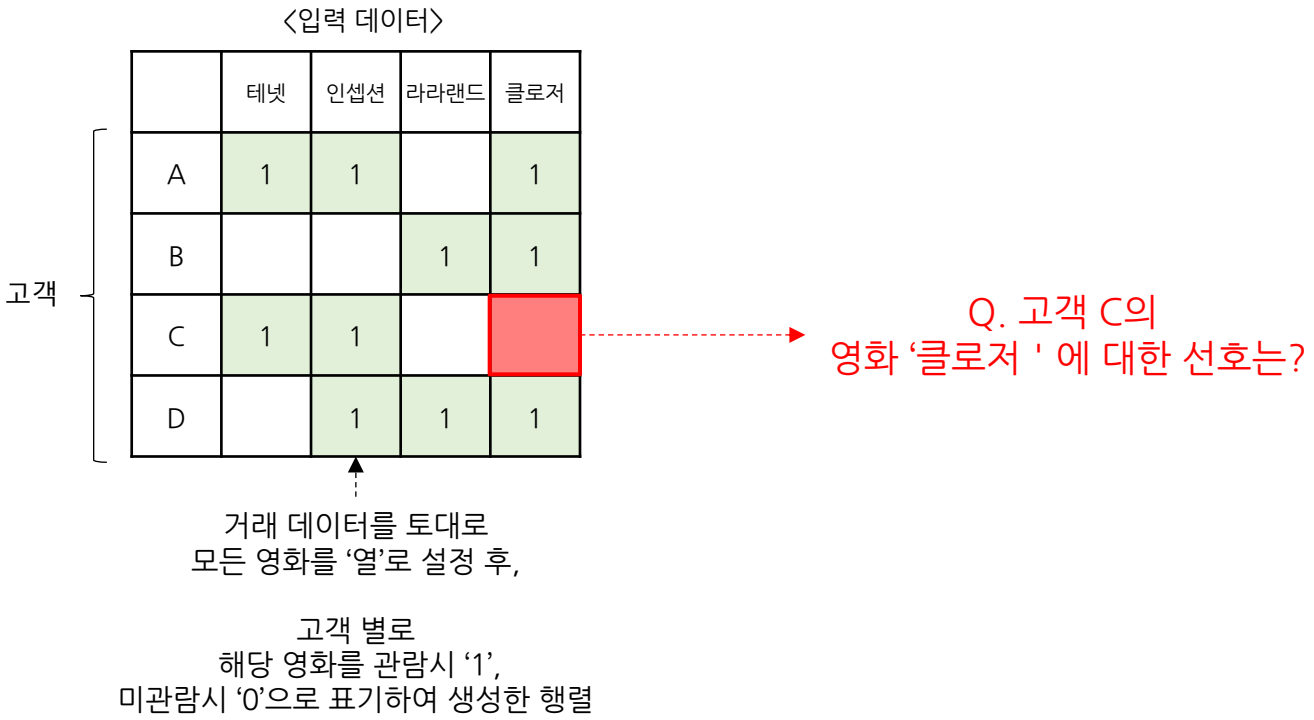
e.g. 위의 예시에서 B는 D그룹, C는 A그룹, '삼진그룹 영어토익반'은 D그룹에 속하게 됩니다.

- 모든 영화들이 매칭 완료
- 최종 클러스터링 결과 : { 영화 A 그룹 : [A, C] } , { 영화 D 그룹 : [D, B, '삼진그룹 영어토익반'] }

- 2) 추천 시스템 : Matrix Factorization

(1) 개요
: 유저와 아이템간의 상호작용 데이터(거래 데이터)를 기반으로 '잠재요인'을 찾아내고,
찾아낸 잠재요인을 기반으로 유저가 관람하지 않은 아이템에 대해서 평가하는 알고리즘

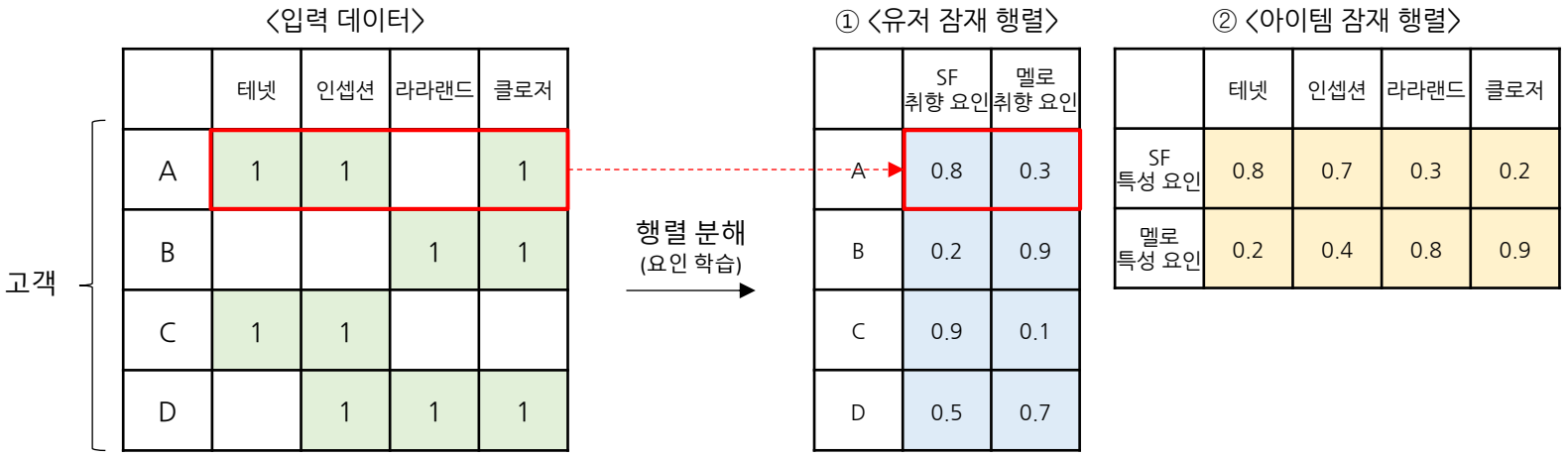
(2) 목표
: 특정 고객의 과거 관람 이력을 토대로 , 아직 관람하지 않은 신규 영화의 선호 예측



- 2) 추천 시스템 : Matrix Factorization

▪ Step ① 유저 잠재 행렬 & 아이템 잠재 행렬 생성

: 주어진 거래데이터(입력 데이터)를 활용해,
유저별로 어떤 *__장르 취향__을 지니고 있는지 & 영화별로는 어떤 *__장르 특성__을 지니고 있는지 학습하고자 합니다.



① 〈유저 잠재 행렬〉
: 위 예시에서 'SF요인'과 '멜로요인' 두 가지로 학습을 진행해, 고객별로 'SF 취향'과 '멜로 취향' 점수를 계산합니다.
e.g. A고객은 [테넷, 인셉션, 클로저] 영화를 관람하였고 그 결과,
"A고객은 'SF 취향' 은 0.8로 높으며, ' 멜로 취향'은 0.3점으로 낮다"고 __고객의 선호__를 정의할 수 있습니다.
또한, 유저 잠재 행렬을 통해서 고객 A와 고객C가 서로 유사한 취향을 갖고 있다고도 알 수 있습니다.

② <아이템 잠재 행렬>
:영화도 동일한 방식으로, 영화별로 'SF 특성'과 '멜로 특성' 점수를 계산하고, 영화의 특성을 정의합니다.

* 실제로는 장르 요인이라고 특정 지을 수 없으나, 편의상 장르라고 이해하시면 됩니다.

- 2) 추천 시스템 : Matrix Factorization

▪ Step ② 예측값 생성

: <유저 잠재 행렬> & <아이템 잠재 행렬>을 토대로,
"고객의 선호"와 "아이템의 특성"을 2가지 요인(SF & 멜로)으로 정의 했습니다.

이는 즉,
우리는 **고객의 선호**를 알고 있고,
동시에 **영화의 특성**을 알고 있다는 의미입니다.

따라서, 고객이 관람하지 않은 영화라고 하더라도,
계산되어 있는 '**고객의 선호**'와 '**영화의 특성**'을 결합하면,
해당 영화를 고객이 얼마나 선호할지 계산 할 수 있습니다.

이제, 우리가 해결하고자 했던, 고객 C의 영화 '클로저'에 대한 선호점수를 계산해봅시다.

<원본 데이터>

	테넷	인셉션	라라랜드	클로저
A	1	1		1
B			1	1
C	1	1		?
D		1	1	1

행렬 분해
(요인 학습)

②<유저 잠재 행렬>

	SF (가정)	멜로 (가상)
A	0.8	0.3
B	0.2	0.9
C	0.9	0.1
D	0.5	0.7

③<아이템 잠재 행렬>

	테넷	인셉션	라라랜드	클로저
SF (가상)	0.8	0.7	0.3	0.2
멜로 (가상)	0.2	0.4	0.8	0.9

* 실제로는 장르 요인이라고 특정 지을 수 없으나, 편의상 장르라고 이해하시면 됩니다.

- 2) 추천 시스템 : Matrix Factorization

▪ Step ② 예측값 생성

	테넷	인셉션	라라랜드	클로저
A	1	1		1
B			1	1
C	1	1		0.27
D		1	1	1

Ex) 유저C의 영화 '클로저' 선호도

	SF (가정)	멜로 (가상)
C	0.9	0.1

X

	클로저
SF (가상)	0.2
멜로 (가상)	0.9

- 0.9(C유저의 SF 선호도) * 0.2(클로저의 SF특성) = 0.18
+
- 0.2(C유저의 멜로 선호도) * 0.9(클로저의 멜로 특성) = 0.09

유저C의 영화 '클로저' 선호도 = 0.27

이처럼 **고객의 선호값**에 **영화의 특성값**을 곱하면, 고객이 관람하지 않은 영화에 대한 선호를 예측할 수 있습니다.
동일한 방법으로 모든 고객의 취향값에 모든 영화의 특성값을 계산하면, 최종 예측 행렬을 얻을 수 있습니다.

	테넷	인셉션	라라랜드	클로저
A	0.7	0.68	0.48	0.43
B	0.34	0.50	0.78	0.85
C	0.74	0.67	0.35	0.27
D	0.54	0.63	0.71	0.73

〈최종 예측 데이터〉

=

	SF 취향 요인	멜로 취향 요인
A	0.8	0.3
B	0.2	0.9
C	0.9	0.1
D	0.5	0.7

〈유저 잠재 행렬〉

X

	테넷	인셉션	라라랜드	클로저
SF 특성 요인	0.8	0.7	0.3	0.2
멜로 특성 요인	0.2	0.4	0.8	0.9

〈아이템 잠재 행렬〉

〈추천 시스템 : Matrix Factorization 참고 사이트〉

- Blog
  - 추천 시스템 - 잠재 요인 협업 필터링
  : https://lsjsj92.tistory.com/564?category=853217
  - 갈아먹는 추천 알고리즘 [3] Matrix Factorization
  : https://yeomko.tistory.com/5?category=805638
  - Matrix Factorization 기술을 이용한 넷플릭스 추천 시스템
  : https://medium.com/curg/matrix-factorization-%EA%B8%B0%EC%88%A0%EC%9D%84-%EC%9D%B4%EC%9A%A9%ED%95%9C-%EB%84%B7%ED%94%8C%EB%A6%AD%EC%8A%A4-%EC%B6%94%EC%B2%9C-%EC%8B%9C%EC%8A%A4%ED%85%9C-7455a40ad527

- Youtube
  - How does Netflix recommend movies? Matrix Factorization
  : https://www.youtube.com/watch?v=ZspR5PZemcs

- 논문
  - Y. Hu et al, Collaborative Filtering for Implicit Feedback Datasets

SPHERE

THANK YOU

김찬웅 대표
Email. chanungkim@tand.kr
Phone. 02-558-8155