콘텐츠 기반 추천 시스템

1팀
20172848 이지평
20172853 장성현
20192761 김정하

AI빅데이터프로젝트
캡스톤 디자인 I

INDEX

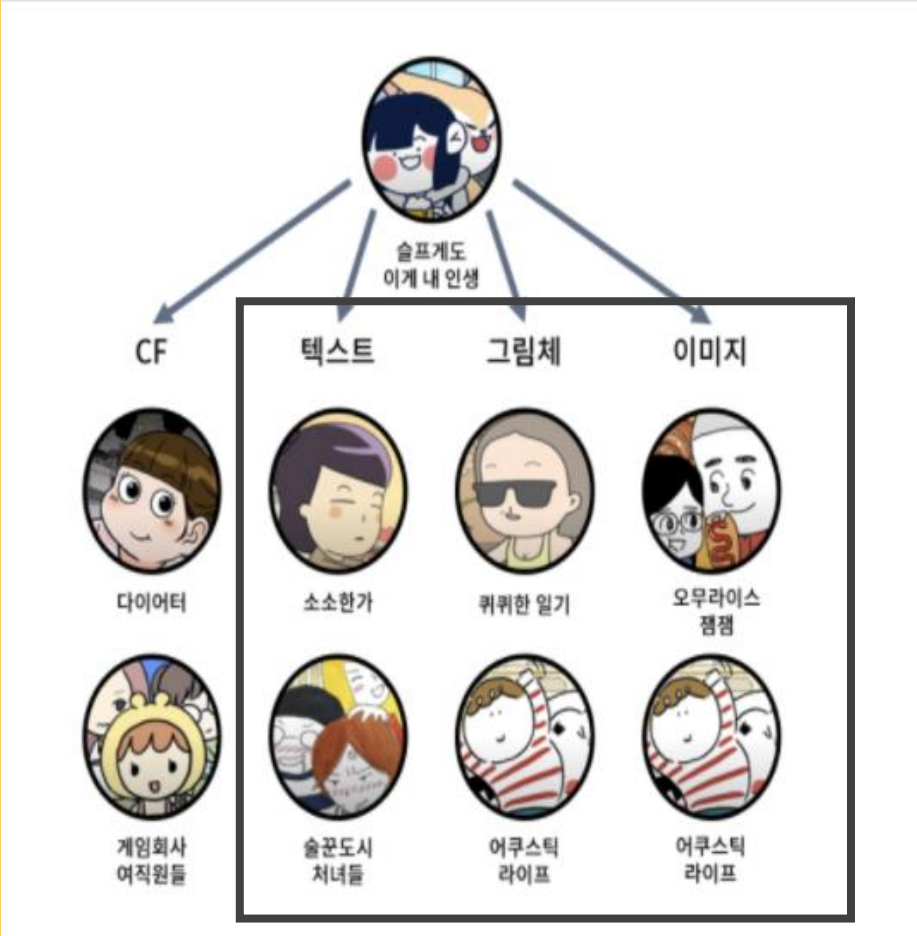
KAKAO WEBTOON

카카오는 어떤 방식으로 웹툰을 추천할까?

해당 작품을 본 사용자가 볼 만한 다른 작품을 추천해 주는 **연관 추천 기능**이 있음

“ ~ 작품과 비슷한 작품들”

→ **콘텐츠 기반 필터링을 활용**



새로운 작품이 추가되면, 콘텐츠 기반 필터링을 위해
아래와 같은 데이터 준비 작업을 실시간으로 진행

- 작품 줄거리 텍스트 데이터를 바탕으로
  작품 **텍스트 임베딩**을 생성

  (텍스트 임베딩에는 카카오가 보유한 다양한 텍스트 데이터로
  사전에 학습한 한국어 임베딩 모델을 사용)

- 카카오 브레인에서 개발한 웹툰 그림체 추출 모델을 사용해
  작품 대표 이미지의 **그림체 임베딩** 생성

- 사전에 다양한 이미지를 사용해 학습 시켜 둔 이미지 모델을
  사용해 작품 대표 **이미지의 임베딩** 생성

01.

콘텐츠 기반 추천시스템

01. 콘텐츠 기반 추천시스템

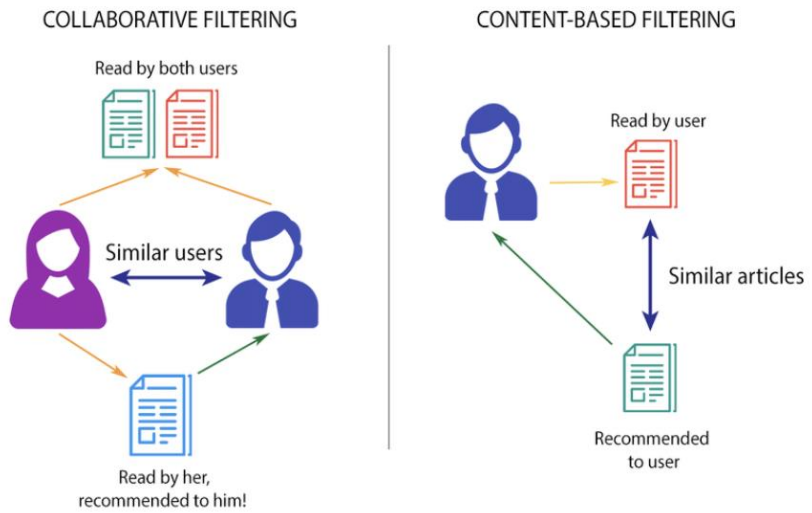
- 콘텐츠 기반 추천시스템이란?

" 사용자가 소비한 아이템에 대해 아이템의 내용(content)이 비슷하거나 특별한 관계가 있는 다른 아이템을 추천하는 방법 "

아이템의 내용은 아이템을 표현할 수 있는 데이터를 지칭
(ex. 아이템 카테고리, 아이템 이름과 같은 텍스트 데이터, 이미지 데이터)

대상 사용자의 평점과 그가 좋아하는 아이템의 속성에 중점을 둠
→ 이를 바탕으로 해당 사용자의 프로파일을 만들어 사용

다른 사용자의 아이템 소비 이력을 활용하는 CF와는
주로 사용하는 데이터가 다르다는 차이점이 있음



[출처 : Designing of Recommendation Engine for Recyclable Waste Mobile App]

01. 콘텐츠 기반 추천시스템

- 콘텐츠 기반 추천시스템

Cold Start 시나리오

- 새 아이템이나 해당 아이템에 대한 평가가 거의 없는 경우
  - → 아이템 속성을 활용하여 추천 가능
- 다른 사용자의 정보가 적은 경우
  - → 자신의 관심사에 관한 충분한 정보를 사용할 수 있다면 추천 가능 (부분적으로 Cold Start 문제 완화)
- 새로운 사용자
  - → 해결 불가능

다만, 다른 사용자의 평점을 사용하지 않기 때문에 추천의 다양성 및 참신성이 떨어질 수 있음

01. 콘텐츠 기반 추천시스템

- 사용하는 데이터

콘텐츠 중심 속성의 다양한 아이템에 대한 설명

명시적 피드백 : 사용자 평점

암시적 피드백 : 사용자 행동

아이템 설명, 사용자 평점, 이와 관련된 분류 및 회귀 모델 결과 등

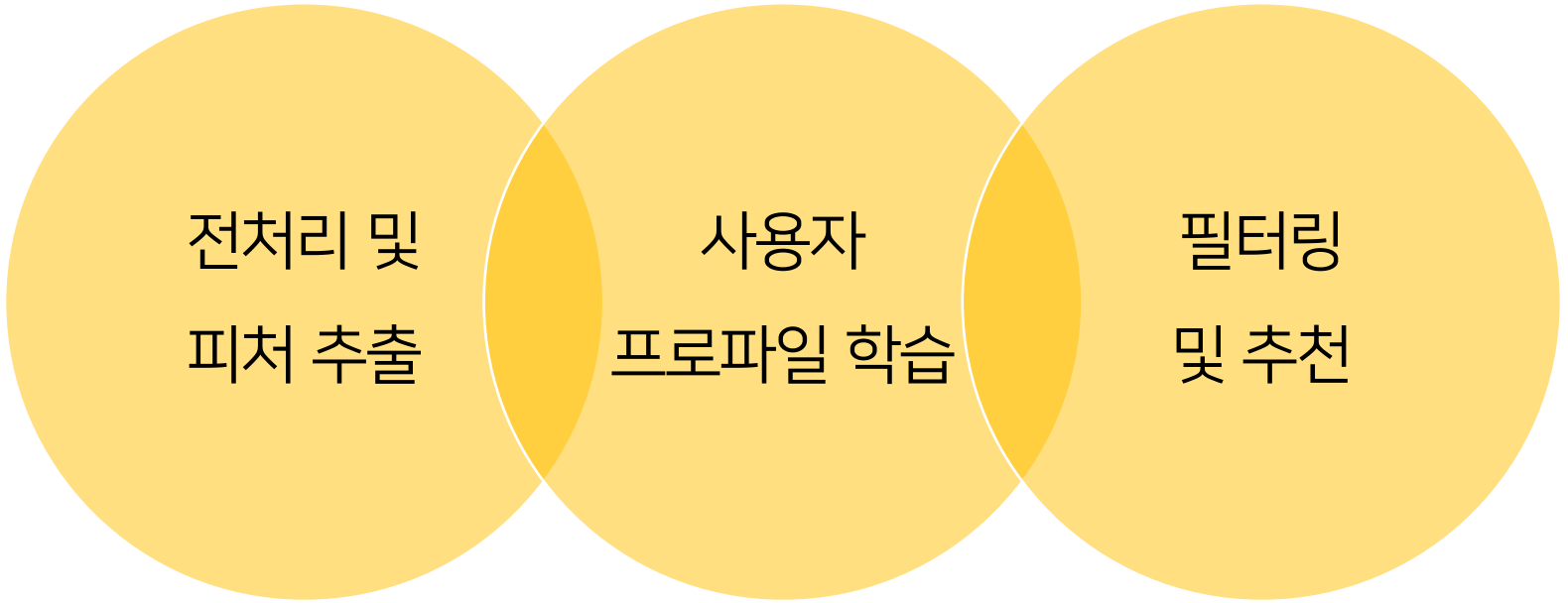
사용자 프로파일

사용자가 관심 키워드에 따라 자신의 프로파일을 설정할 수도 있음

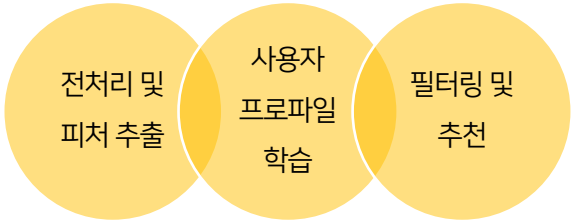
02.

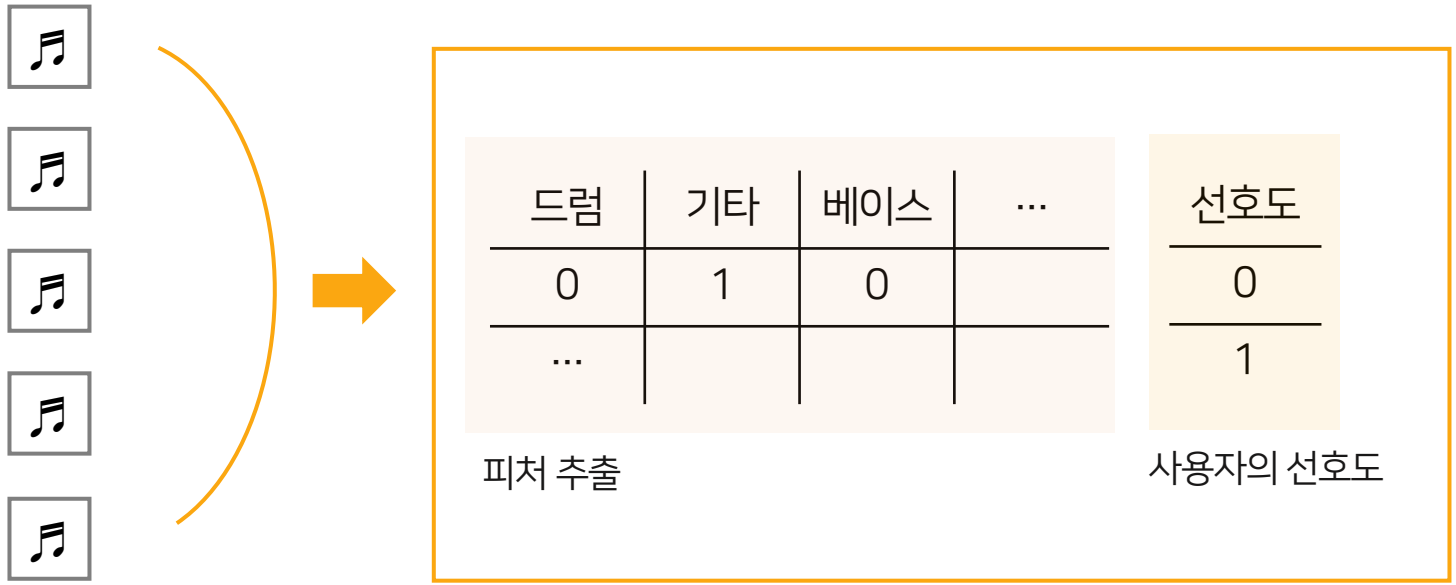
콘텐츠 기반 추천시스템
구성요소

02. 콘텐츠 기반 추천시스템의 구성요소



02. 콘텐츠 기반 추천시스템의 구성요소





학습

필터링 및 추천

사용자 프로파일

02. 콘텐츠 기반 추천시스템의 구성요소

전처리 및 피처 추출

- 피처 추출

아이템 간의 차이를 분간할 수 있는 피처 추출

일반적으로 기본 데이터에서 키워드 추출

다양한 소스(필드)에서 피처를 추출 → 전처리 → **키워드 기반 벡터 공간으로 표현**

분류 과정에서 쉽게 사용하기 위해 피처 선택 및 가중치 설정 필요 (지도, 비지도 방식 존재)

사용 중인 애플리케이션에 따라 추출할 피처 유형의 변화가 큼

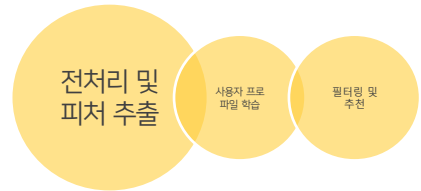
02. 콘텐츠 기반 추천시스템의 구성요소

전처리 및 피처 추출

- 사용자의 선호도에 대한 정보 수집

- 명시적 피드백 : 평점
- 암시적 피드백 : 주로 긍정적인 선호도만 포착 가능
- 텍스트 의견
- 사례
- 최근접 이웃 or 로키오 분류를 통해 암시적 피드백으로 사용

다만, 콘텐츠 기반 추천시스템과 지식 기반 추천시스템의 경계 모호



02. 콘텐츠 기반 추천시스템의 구성요소

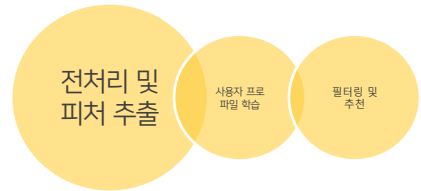
전처리 및 피처 추출

- 피처 표현 및 정제

텍스트 도메인에서 모델 적용 전 필요한 전처리 진행

1. 불용어 제거 : 통상 관사, 전치사, 접속사 및 대명사 등 아이템과 관련이 없는 불용어 제거
2. 형태소 분석 : 동일한 단어의 유사어 통합 (단수, 복수, 시제 통합)
3. 구문 추출 : 문서에서 함께 자주 발생하는 단어 검색→ 벡터 공간 표현으로 변환

   - One-hot-encoding
   - TF-IDF
   - Word2Vec



02. 콘텐츠 기반 추천시스템의 구성요소

전처리 및 피처 추출_피처 선택 및 가중치 설정

가장 유익한 단어만 벡터 공간 표현에 남도록 키워드의 개수에 대해 크기 제한 필요

노이즈가 많은 단어는 과적합을 만들기 때문에 우선적으로 제거

비지도 방식이 아닌 지도 방식의 피처 선택 (라벨 : 사용자의 피드백)

ex. 지니계수, 엔트로피, χ^2 통계, 정규화된 편차

02. 콘텐츠 기반 추천시스템의 구성요소

전처리 및 피처 추출_피처 선택

- 지니 계수

피처 선택에서 가장 일반적으로 사용하는 방법 중 하나

사용 가능한 평점의 총 수 : $p_1(w), \dots, p_t(w)$ 는 t값에서 평가된 아이템의 비율

단어 w의 지니 인덱스 : $Gini(w) = 1 - \sum_{i=1}^{t} p_i(w)^2$

지니(w) 값 범위 : (0, 1-1/t)

작은 값 → 더 큰 식별력

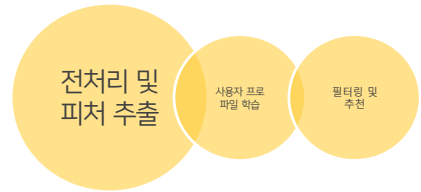
- 엔트로피

지니 계수와 유사

엔트로피$(w) = -\sum_{i=1}^{t} p_i(w)\log(p_i(w))$

엔트로피 값 범위 : (0, 1)

작은 값 → 더 큰 식별력



02. 콘텐츠 기반 추천시스템의 구성요소

전처리 및 피처 추출_피처 선택

- χ^2-통계

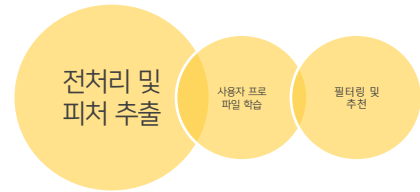
단어와 클래스 사이의 동시 발생을 분할표로 처리해 계산
분할표의 다양한 셀에서 관찰 값과 예상 값 사이의 정규화된 편차를 측정

O_i : i번째 셀의 관측 값 / E_i : i번째 셀의 예상 값

$$\chi^2 = \sum_{i=1}^{P} \frac{(O_i - E_i)^2}{E_i} \qquad (4.5)$$

명시적으로 기대값 계산 없이 분할표에서 관찰한 값의 함수로 χ^2 통계량 계산 가능
(기대값이 행과 열의 관측값을 집계한 함수이기 때문에)

$$\chi^2 = \frac{(O_1+O_2+O_3+O_4)*(O_1O_4-O_2O_3)^2}{(O_1+O_2)*(O_3+O_4)*(O_1+O_3)*(O_2+O_4)} \qquad (4.6)$$



02. 콘텐츠 기반 추천시스템의 구성요소

전처리 및 피처 추출_피처 선택

- χ^2 통계

[기대값 분할표]

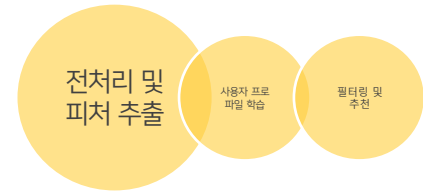
	설명에 단어 출현	설명에 단어 미출현
아이템 구매 시	1000*0.1*0.2 = 20	1000*0.1*0.8=80
아이템 미구매 시	1000*0.9*0.2 = 180	1000*0.9*0.8=720

[관찰값 분할표]

	설명에 단어 출현	설명에 단어 미출현
아이템 구매 시	O_1 = 60	O_2 = 40
아이템 미구매 시	O_3 = 140	O_4 = 760

각각 수식 4.5 / 4.6 계산을 진행하면 동일한 값을 얻을 수 있음
→ 명시적으로 기대값 계산 없이 분할표에서 관찰한 값의 함수로 χ^2 통계량 계산 가능

χ^2 통계의 값이 클수록 특정 용어와 아이템이 더 큰 관련이 있음



02. 콘텐츠 기반 추천시스템의 구성요소

전처리 및 피처 추출_피처 선택

- 정규화된 편차
: 평점의 상대적 순서를 잃어버리지 않는 방식

σ^2 : 평점의 분산
$\mu^+(w)$: 단어 w가 포함된 모든 문서의 평균 평점
$\mu^-(w)$: 단어 w를 포함하지 않는 모든 문서의 평균 평점

$$Dev(w) = \frac{|\mu^+(w) - \mu^-(w)|}{\sigma}$$

$Dev(w)$의 값이 클수록 더 차별적인 단어 나타냄

이러한 접근법은 평점이 수치적 종속 변수일때 (카테고리 종속변수일 때는 Fisher의 차별 지수 사용 권장)

02. 콘텐츠 기반 추천시스템의 구성요소

전처리 및 피처 추출_피처 가중치 설정

- 지도 방식의 피처 가중치 설정

피처 선택 측정값을 가져와 가중치 유도

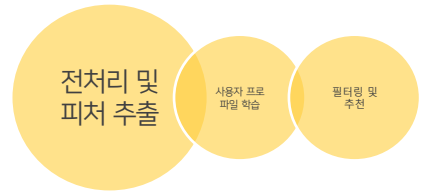
Ex) $g(x) = a - Gini(w)$ 이때 $g(x)$는 $(a-1,\ a)$ 의 범위를 가짐

a 의 값이 작으면 → 민감도 높아짐

a 는 가중 함수의 매개변수로 볼 수 있음

각 단어 w 의 가중치는 $g(w)$로 곱함

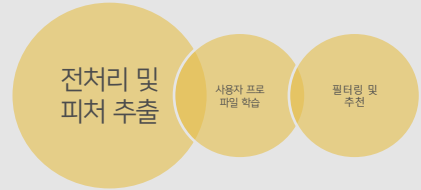
피처 가중치 설정은 피처 선택보다 더 소프트한 방식



02. 콘텐츠 기반 추천시스템의 구성요소

학습

필터링 및 추천



02. 콘텐츠 기반 추천시스템의 구성요소

사용자 프로파일 학습

Ratings matrix에서 rating을 개별 값(좋아요 or 싫어요)으로 처리하면 분류 task
수치 값으로 처리하면 회귀 task

학습 문서 : 전처리 및 피처 선택 단계에서 추출된 아이템 설명 + 활성 사용자가 학습 문서에 지정한 평가
학습 모델 : 특정 사용자에게만 적용.

텍스트 도메인에서의 분류 및 회귀 모델링과 유사함.

$\boldsymbol{D_L}$: 학습 문서 세트

$\boldsymbol{D_U}$: 테스트 문서(잠재적으로 사용자에게 추천될 수 있지만 사용자가 아직 미구매 하거나 평가하지 않은 아이템 설명)

→ $\boldsymbol{D_L}$ 의 학습모델은 $\boldsymbol{D_U}$에서 현재 사용자에게 추천을 제공하는 데 사용

02. 콘텐츠 기반 추천시스템의 구성요소

02. 콘텐츠 기반 추천시스템의 구성요소

사용자 프로파일 학습 : 최근접 이웃 분류 (KNN)

가장 단순한 분류 기술 중 하나. 비교적 직접적인 방법으로 구현 가능

첫번째 단계 : 유사함수 정의

- 수치적 속성 : 코사인 유사도, 유클리드 거리, 맨해튼 거리 같은 유사도/거리 함수
- 범주적 속성 : 일치 기반 유사도 측정 값

유사도함수 사용함으로써 얻는 효과

1. 사용자 선호도가 알려지지 않은 아이템(문서) 예측 유용

   (수치형)각 아이템의 k이웃에 대한 평점의 평균 값 결정 → 해당 아이템에 대한 예상 평점

   (범주형)평점의 각 값에 대한 투표수 결정 → 가장 높은 빈도를 가진 평점 값으로 예측

2. 유사도 값으로 각 평점에 가중치 부여 가능

→ 최근접 이웃 결정 / 각각의 최근접 이웃 결정에 필요한 시간은 D_L 크기에 선형적임.

→ 높은 계산 복잡도

02. 콘텐츠 기반 추천시스템의 구성요소

사용자 프로파일 학습 : 최근접 이웃 분류 (KNN)

- 높은 계산 복잡도 줄이는 방법

  “ 클러스터링을 사용해 D_L 의 학습 문서를 줄이는 것 “

  → 대상 문서와 비교적 적은 수의 집계 문서 사이의 유사도를 계산하므로 분류 프로세스의 속도를 높임

  → 클러스터링에 대한 전처리 추가 발생 시간 < 클러스터링 적용 전 시간

02. 콘텐츠 기반 추천시스템의 구성요소

사용자 프로파일 학습 : 최근접 이웃 분류 (KNN)

- 최근접 이웃 분류와 사례 기반 추천시스템

- 최근접 이웃 방법은 일반적으로 지식 기반 추천시스템, 특히 사례 기반 추천시스템과 연결됨

주요 차이점

- 사례 기반 추천 시스템은 사용자가 대화형으로 관심있는 한 개의 사례를 선택하면 사용자가 관심 갖는 해당 사례의 가능한 아이템들을 받을 수 있음, 오직 하나의 사례만 이용할 수 있으므로 유사도 함수를 설계할 때 상당한 양의 도메인 지식을 사용
- 지식 기반 시스템에서는 과거 데이터 또는 평점을 덜 이용

02. 콘텐츠 기반 추천시스템의 구성요소

사용자 프로파일 학습 : 베이즈 분류 모델

$c_i \in \{-1, 1\}$ 긍정부정 이진 분류

레이블이 붙은 집합 → 사용자 프로파일

베르누이 모델에서 단어의 빈도는 무시되며 문서에서 단어의 존재 여부만 고려

→ 0과 1의 값만을 포함하는 d개 단어의 이진 벡터로 취급

02. 콘텐츠 기반 추천시스템의 구성요소

사용자 프로파일 학습 : 베이즈 분류 모델

$\bar{X}$: D_U 에 포함된 대상문서

$\{x_1, x_2 \ \ldots x_d\}$: $\bar{X}$ 에 있는 d개의 이진 특성,

x_i는 i번째 단어가 문서 $\bar{X}$ 에 존재하는지 여부에 따라 0-1값

$c(\bar{X})$: $\bar{X}$의 클래스(이진 평점) → $P(c(\bar{X})=1 \mid x_1, x_2, \ldots, x_d)$ 와 $P(c(\bar{X})=-1 \mid x_1, x_2, \ldots x_d)$를 정하고 둘 중 큰 것 선택

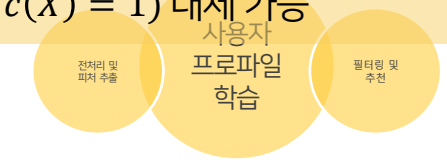
합 = 1

이는 활성 사용자가 베이즈 규칙을 사용해 평가할 수 있으며 나이브 가정을 적용해 평가할 수 있음

$$P(c(\bar{X})=1 \mid x_1, x_2 \ \ldots x_d) = \frac{P(c(\bar{X})=1)P(x_1,,,x_d|c(\bar{X})=1)}{P(x_1,,,x_d)} \propto P(c(\bar{X})=1)P(x_1,,,x_d|c(\bar{X})=1)$$

$$= P(c(\bar{X})=1)\prod_{i=1}^{d} P(x_i|c(\bar{X})=1)$$

나이브 가정에서 문서에서의 단어 발생은 조건부이므로 독립된 사건
→ $P(x_1,,,x_d|c(\bar{X})=1)$를 $\prod_{i=1}^{d} P(x_i|c(\bar{X})=1)$ 대체 가능



02. 콘텐츠 기반 추천시스템의 구성요소

사용자 프로파일 학습 : 베이즈 분류 모델

$$\frac{P(c(\bar{X})=1)P(x_1,,,x_d|c(\bar{X})=1)}{P(x_1,,,x_d)} \Rightarrow K \propto P(c(\bar{X})=1)P(x_1,,,x_d|c(\bar{X})=1)$$

분모가 클래스와 독립적(클래스의 상대적인 순서(=한 아이템에서 평점별 확률)에 영향 X) → $\propto$ 사용 가능

But 분모는 사용자 선호 아이템 성향의 순위 측면에서 중요한 역할

→ 정보의 손실을 완화하기 위해 비례상수(K) 사용

앞서 말했 듯 $c(\bar{X})$의 모든 확률(1, -1)의 합이 1이기 때문에 K에 대해 다음 값을 유도할 수 있음

$$K\left[P(c(\bar{X})=1) \prod_{i=1}^{d} P(x_i \mid c(\bar{X})=1) + P(c(\bar{X})=-1) \prod_{i=1}^{d} P(x_i \mid c(\bar{X})=-1) \right] = 1$$

$$K = \frac{1}{P(c(\bar{X})=1) \prod_{i=1}^{d} P(x_i \mid c(\bar{X})=1) + P(c(\bar{X})=-1) \prod_{i=1}^{d} P(x_i \mid c(\bar{X})=-1)}$$

02. 콘텐츠 기반 추천시스템의 구성요소

사용자 프로파일 학습 : 베이즈 분류 모델 (중간 확률 추정)

$P(c(\bar{X})=1|x_1, x_2, \ldots, x_d) \propto P(c(\bar{X}) = 1) \prod_{i=1}^{d} P(x_i|c(\bar{X}) = 1)$

$P(c(\bar{X})=-1|x_1, x_2, \ldots, x_d) \propto P(c(\bar{X}) = -1) \prod_{i=1}^{d} P(x_i|c(\bar{X}) = -1)$

이를 계산하려면 오른쪽의 확률을 추정 해야함

이전 클래스 확률($P(c(\bar{X})=1)$, $P(c(\bar{X})=-1)$), 피처 별 조건부 확률($P(x_i|c(\bar{X}) = 1)$, $P(x_i|c(\bar{X}) = -1)$)을 추정해야 함

$P(c(\bar{X}))=1$ 는 레이블링 된 데이터 D_L에서 긍정 학습 예제 D_L^+ 의 비율로 추정할 수 있음

02. 콘텐츠 기반 추천시스템의 구성요소

사용자 프로파일 학습 : 베이즈 분류 모델 (중간 확률 추정)

과적합을 줄이기 위해 작은 매개변수 α (> 0) 에 비례하는 값을 분자와 분모에 더하는 라플라시안 평활화 수행

$$P(c(\bar{X}) = 1) = \frac{|D_L{}^+| + \alpha}{|D_L| + 2\alpha}$$

$P(x_i \mid C(\bar{X}) = 1)$: i번째 피처가 x_i의 값을 취하는 긍정 클래스의 인스턴스의 일부로 추정

$q^+(x_i)$: i번째 피처에 대해 x_i의 값을 취하는 긍정 클래스의 인스턴스 수

다음 라플라시안 평활화 매개변수 β (> 0)를 사용해 추정 가능

$$\mathrm{P}(\mathrm{x_i}|\mathrm{c}(\bar{\mathrm{X}}) = 1) = \frac{\mathrm{q}^+(\mathrm{x_i}) + \beta}{|\mathrm{D_L}{}^+| + 2\beta}$$

라플라시안 평활화는 학습 데이터가 거의 없는 경우(양이 제한됐을 때)에 유용

ex) D_L^+가 비어있는 경우(좋아요가 없는 경우) $\mathrm{P}(\mathrm{x_i}|\mathrm{c}(\bar{\mathrm{X}}) = 1)$ 의 확률은 사전에 가지고 있던 믿음으로 0.5로 추정

평활화가 없다면 분자와 분모가 모두 0이 될 것

02. 콘텐츠 기반 추천시스템의 구성요소

사용자 프로파일 학습 : 베이즈 분류 모델의 예

키워드 → / 노래-idf ↓	드럼	기타	비트	클래식	교향곡	오케스트라	좋아요 또는 싫어요
1	1	1	1	0	0	0	싫어요
2	1	1	0	0	0	1	싫어요
3	0	1	1	0	0	0	싫어요
4	0	0	0	1	1	1	좋아요
5	0	1	0	1	0	1	좋아요
6	0	0	0	1	1	0	좋아요
테스트-1	0	0	0	1	0	0	?
테스트-2	1	0	1	0	0	0	?

열: 노래의 속성을 나타내는 피처(마지막 열: 좋음/싫음→ 평가)
행: 사용자 프로파일(마지막 2행: 평가해야 할 후보 음악 → 테스트 인스턴스)

$$-\ P(\text{좋아요}|\text{Test}-1) \propto 0.5\prod_{i=1}^{6} P(\text{좋아요}|x_i) = (0.5) * \frac{3}{4} * \frac{2}{2} * \frac{3}{4} * \frac{3}{3} * \frac{1}{4} * \frac{1}{3} = \frac{3}{128}$$

$$-\ P(\text{싫어요}|\text{Test}-1) \propto 0.5\prod_{i=1}^{6} P(\text{싫어요}|x_i) = (0.5) * \frac{1}{4} * \frac{0}{2} * \frac{1}{4} * \frac{0}{3} * \frac{3}{4} * \frac{2}{3} = 0$$

$P(\text{좋아요}|\text{Test}-1)$ 이 1이고 $P(\text{싫어요}|\text{Test}-1)$ 가 0의 결과,
test-2는 반대의 결과 → test-1이 추천

라플라시안 평활화를 사용할 때, 클래스 중 하나가 다른 클래스보다 훨씬 높은 확률을 얻지만,
다양한 클래스에 대해 이진 확률 값을 얻지 못할 것
→ 예상되는 "좋아요"의 확률로 순위를 매겨 사용자에게 추천!

02. 콘텐츠 기반 추천시스템의 구성요소

사용자 프로파일 학습 : 규칙 기반 분류 모델

규칙의 지지도 : 선행 규칙(A)과 결과 규칙(B)을 모두 충족시키는 행의 비율 - $\boldsymbol{P(A \cap B)}$

규칙의 신뢰도 : 선행을 만족시키는 것으로 이미 알려진 행에서 결과를 만족시키는 행의 비율 - $\boldsymbol{P(B|A)}$

- 협업 필터링의 규칙 기반 분류와 유사 : 협업 필터링의 아이템-아이템 규칙에서, 선행 항목과 결과 규칙은 아이템의 평점에 해당
- 차이점 : 협업 필터링에서 규칙의 선행은 다양한 아이템의 평가에 해당하는 반면, 콘텐츠 기반 방법에서는 규칙의 선행 아이템 설명에서

  특정 키워드의 존재와 일치한다는 것

  Ex) 아이템이 키워드 세트 A를 포함 -> 평가 -> 좋아요

  아이템이 키워드 세트 B를 포함 -> 평가 -> 싫어요

첫번째 단계 : 활성 사용자 프로파일을 활용해 원하는 규칙에 따라 모든 규칙을 지원

→ 현재 활동 중인 사용자에게만 적용

02. 콘텐츠 기반 추천시스템의 구성요소

사용자 프로파일 학습 : 규칙 기반 분류 모델의 예

접근 방식

1. (학습 단계) 학습 데이터 세트 D_L 에서 원하는 수준의 **최소 지지도 및 신뢰도 수준**에서 사용자 프로파일의 모든 관련 규칙을 결정한다.

2. (테스
결정

→ 높은

키워드 → / 노래-idf↓	드럼	기타	비트	클래식	교향곡	오케스트라	좋아요 또는 싫어요
1	1	1	1	0	0	0	싫어요
2	1	1	0	0	0	1	싫어요
3	0	1	1	0	0	0	싫어요
4	0	0	0	1	1	1	좋아요
5	0	1	0	1	0	1	좋아요
6	0	0	0	1	1	0	좋아요
테스트-1	0	0	0	1	0	0	?
테스트-2	1	0	1	0	0	0	?

예시) 33% 지지도 수준과 75% 신뢰도 수준에서 지지도 규칙 값과 함께 다음 규칙 생성

규칙 1 : {클래식} -> 좋아요(50%, 100%)

규칙 2 : {교향곡} -> 좋아요(33%, 100%)

규칙 3 : {클래식, 교향곡} -> 좋아요(33%, 100%)

규칙 4 : {드럼, 기타} -> 싫어요(33%, 100%)

규칙 5 : {드럼} -> 싫어요(33%, 100%)

규칙 6 : {비트} -> 싫어요(33%, 100%)

규칙 7 : {기타} -> 싫어요(50%, 75%)

주로 신뢰도가 떨어지는 순서로 정렬, 지지도가 감소하는 순서로 연결이 끊어짐.

규칙 1은 테스트-1에 의해 실행되는 반면, 규칙 5와 6은 테스트-2에 의해 실행됨.

따라서 테스트-2보다 테스트-1이 활성 사용자에게 추천되어야 함.

이러한 설명은 종종 고객의 관점과 판매자의 관점 모두에서 추천 시스템에 매우 유용.

02. 콘텐츠 기반 추천시스템의 구성요소

사용자 프로파일 학습 : 회귀 기반 분류 모델

바이너리 평가, 구간 기반 평가, 수치 평가 같은 다양한 유형의 평가에 사용할 수 있음
선형모델, 로지스틱 회귀모델 및 순서형 프로빗 모형과 같은 대규모 회귀 모델 사용 → 평점 모델링

D_L : 크기 d의 어휘집에 대해 레이블링된 학습 세트 D_L의 n개의 문서를 나타내는 n x d 행렬

$\bar{y}$: 학습세트의 n개의 문서에 대한 활성 사용자의 평가를 포함하는 n차원 열 벡터

$\bar{W}$: 선형함수의 각 단어의 계수를 나타내는 d-차원 행 벡터

O : 목적함수 → 이것을 최소화하는 $\bar{W}$ 를 구해야 함

02. 콘텐츠 기반 추천시스템의 구성요소

사용자 프로파일 학습 : 회귀 기반 분류 모델

D_L : 크기 d의 어휘집에 대해 레이블링된 학습 세트 D_L의 n개의 문서를 나타내는 n x d 행렬
$\bar{y}$: 학습세트의 n개의 문서에 대한 활성 사용자의 평가를 포함하는 n차원 열 벡터
$\bar{W}$: 선형함수의 각 단어의 계수를 나타내는 d-차원 행 벡터
O : 목적함수 → 이것을 최소화하는 $\bar{W}$ 를 구해야 함

$$\bar{y} \approx D_L \bar{W}^T$$

$$Minimize\ O = \|D_L \bar{W}^T - \bar{y}\|^2 + (\lambda \|\bar{W}\|^2) \rightarrow \text{정규화}$$

$$D_L{}^T (D_L \bar{W}^T - \bar{y}) + \lambda \bar{W}^T = 0$$

$$(D_L{}^T D_L + \lambda I) \bar{W}^T = D_L{}^T \bar{y}$$

$$\bar{W}^T = (D_L{}^T D_L + \lambda I)^{-1} D_L{}^T \bar{y}$$

레이블이 지정되지 않은 집합 D_U 로부터 임의의 주어진 문서 벡터(아이템 설명) $\bar{X}$의 평가는 $\bar{W}$와 $\bar{X}$ 사이의 내적으로 예측

02. 콘텐츠 기반 추천시스템의 구성요소

사용자 프로파일 학습 : 회귀 기반 분류 모델

회귀 모델	평가의 성격(목표변수)
선형 회귀	실수
다항 회귀	실수
커널 회귀	실수
이진 로지스틱 회귀	단항, 이진
다변 로지스틱 회귀	범주형, 순서형
프로빗	단항, 이진
다변 프로빗	범주형, 순서형
순서형 프로빗	순서형, 구간 기반

회귀모델 계열 및 다양한 유형의 평가에 대한 적용 가능성

03.

콘텐츠 기반 추천시스템과 협업 필터링

03. 콘텐츠 기반 추천시스템과 협업 필터링

장점

- 주어진 사용자에 의해 평가된 이전 아이템들이 추천을 만들기 위해 활용
  신규 사용자가 아닌 이상, 항상 의미있는 추천 생성 가능
  콘텐츠 기반 시스템은 새로운 사용자에 대해서만 콜드 스타트 문제 있음
- 콘텐츠 기반 방법은 아이템의 피처 측면에서 설명 제공
- 일반적으로 상용 텍스트 분류 모델과 함께 사용 가능. 협업 시스템과 달리 각 사용자별 분류는 큰 문제 아님.
  상대적으로 적은 엔지니어링 노력으로 쉽게 사용

단점

- 사용자가 지금까지 본 것과 유사한 아이템을 찾는 경향(과적합)
  일정 수준의 참신함(novelty)과 의외성(serendipity)을 가지는 것이 바람직하나 콘텐츠 기반 시스템은 그러지 못함
  과도한 특수화와 뜻밖의 발견을 하기 부족한 점은 콘텐츠 기반 추천 시스템의 가장 중요한 두 가지 과제임
- 새로운 아이템에 대한 콜드 스타트 문제를 해결하는 데 도움이 되지만
  신규 사용자를 위해 이러한 과적합 문제를 해결하는 데에는 도움이 되지 않음

→ 콘텐츠 기반 시스템은 거의 별개로 사용하지 않으며 다른 유형의 추천 시스템과 함께 사용

03. 콘텐츠 기반 추천시스템과 협업 필터링

협업 필터링 시스템을 위한 콘텐츠 기반 모델 사용

- 콘텐츠 기반 방법을 협업 필터링에서 직접 사용 가능
- 아이템의 콘텐츠 설명은 설명 키워드 참조하지만, 사용자의 평점을 활용해 콘텐츠 기반 설명을 정의하는 시나리오 계획 가능

  → 각 아이템에 대해 아이템을 평가한 사용자의 사용자 이름을 이 평가의 값과 연결해 새 ' 키워드 ' 를 만듦.

  ex) 터미네이터 : 존#좋아요, 앨리스#싫어요, 톰#좋아요

- 가능한 평가 수가 적은 경우 더욱 효과적임
- 적절한 콘텐츠 표현(키워드)을 정의하고 기성 콘텐츠 기반 방법을 직접 사용함으로써 협업 필터링에 대한 많은 방법 포착 가능
- 이러한 접근 방식은 더 이상 다른 사용자의 사용 가능한 평점 데이터를 낭비하지 않으며 통합 프레임워크 내에서 콘텐츠 기반 및 협업 모델의 힘을 결합함

03. 콘텐츠 기반 추천시스템과 협업 필터링

사용자 프로파일 활용

- 콘텐츠 속성으로 협업 필터링과 같은 모델을 작성할 수 있는 또 다른 경우 : 사용자 프로파일이 지정된 키워드의 형태로 사용 가능할 때

ex) 사용자는 특정 관심사를 키워드 형태로 지정할 수 있음

→ 사용자 피처를 사용해 모든 사용자에 대해 전역 분류 모델 생성 가능

→ 각 사용자 및 아이템의 속성 벡터의 크로네커 곱을 사용해 콘텐츠 중심 표현 만들 수 있음

→ 사용자-아이템 조합 평점으로 매핑하기 위해 이 표현에 분류 또는 회귀 모델 구성

THANK YOU

콘텐츠 기반 추천시스템

1팀
20172848 이지평
20172853 장성현
20192761 김정하

AI빅데이터 프로젝트 : 캡스톤 디자인 I