4. 콘텐츠 기반 추천 시스템

추천시스템 4주차

4.1 개요

- 협업필터링 (CF) 추천 시스템
  - 평점 행렬 내 상관 관계를 이용하여 추천
  - 유사도
    - 피어그룹 내 상관 관계에 기반
- 콘텐츠 기반 (Content-Based) 추천 시스템
  - 아이템에 포함된 설명력이 있는 속성을 활용
  - 사용자 자신의 평점과 아이템에 대한 상호작용을 활용
  - 신규 아이템 또는 평점이 드문 경우에 유용
  - 유사도
    - 사용자가 좋아하는 아이템의 속성 기반 (다른 사용자 역할이 거의 없음)

4.1 개요

- 콘텐츠 기반 시스템의 데이터 소스
  - 아이템에 대한 콘텐츠 성격의 설명
    - 예: 상품에 대한 설명서
  - 피드백을 가공한 사용자 프로파일
    - 예: 사용자가 평점을 매긴 아이템 설명 문서 / 또는 이를 활용한 분류 모델
      - 설명 문서: 피쳐
      - 평점: 타겟 (레이블)
    - 예: 사용자가 직접 관심 키워드를 입력한 내용 (지식 기반 추천 시스템과 유사)
- 다른 사용자의 평점을 이용하지 않음
  - 장점: 콜드 스타트 문제
    - 신규 아이템이라도, 사용자 평점 정보가 어느 정도 있다면 평점 예측
    - 신규 사용자에 대한 추천은 여전히 어려움
  - 단점: 다양성 (diversity), 참신성 (novelty) 감소
    - 유사한 속성을 가진 아이템 위주의 추천이기 때문

4.1 개요

- 콘텐츠 기반 시스템이 유망한 영역
  - 아이템의 속성 정보가 매우 많은 경우
  - 풍부한 텍스트 / 비구조화된 도메인
    - 이커머스에서의 상품 설명, 웹페이지
- 하이브리드 추천 시스템
  - 협업필터링, 콘텐츠기반, 지식기반 시스템들을 결합
- 4장 구성
  - 콘텐츠 기반 시스템 기본 구성 요소
  - 피처 추출
  - 사용자 프로파일릴 학습및 추천
  - 시스템 비교 및 융합

4.2 콘텐츠 기반 시스템의 기본 구성 요소

- 전처리 및 피처(속성) 추출
  - 다양한 소스로부터 피처를 추출
    - 웹페이지, 제품 설명, 뉴스, (음악, 이미지 등)
    - 벡터 공간으로 표현
      - 키워드
      - 이미지 등
- 사용자 프로파일의 콘텐츠 기반 학습
  - 개별 사용자의 아이템 관심사를 예측하는 모델
    - 사용자 피드백 (평점, 인터랙션)
    - 상품 피처를 결합
  - ‘모델’을 곧 ‘사용자 프로파일 ' 이라 할 수 있음
- 필터링 및 추천
  - 모델을 활용하여 개별 사용자에 아이템 추천

4.3.1 피처 추출

- 변별력이 있는 피처 (Discriminative feature)
  - 사용자 관심사의 예측력에 큰 기여를 하는 피처
- 피처 추출
  - 아이템에 대한 설명 (description)을 추출
    - 키워드, (이미지, 카테고리 등의 다양한 메타 정보)
    - 예) 책
      - 요약문
      - 대표 키워드
      - 타이틀
      - 저자
      - 가격대
      - 하드커버, 소프트커버
      - 등

4.3.1.1 영화 추천의 예

- IMDb: 영화 추천
  - 각 영화
    - 시놉시스, 감독, 배우, 장르 등의 설명 (description) + 사용자태그들
    - “Shrek“

      “After his swamp is filled with magical creatures, an ogre agrees to rescue a princess for a villainous lord in order to get his land back.”

- 특징
  - 각 키워드가 추천에 기여하는 중요도가 동일하지 않음
    - 도메인 지식 활용: ‘제목’과 ‘주연배우’에 대한 키워드가 훨씬 중요할 수 있음을 반영
    - 피처간의 상대적인 중요도를 측정할 수 있는 자동화된 방법 적용

4.3.2 텍스트: 피처 표현 및 정제 (Feature Representation and Cleaning)

- 키워드 모음 (Bag of words) 결정
  - 불용어 제거 (stop-word removal)
    - “a”, “an”, “the” 등 고빈도 키워드를 삭제
  - 형태소 분석 (Stemming)
    - 동일 단어의 변형을 통일
    - “hoping”, “hope” -> “hop”
  - 구 추출 (Phrase extraction)
    - 동시에 자주 등장하는 구를 결정
    - “hot dog”

4.3.2 텍스트: 피처 표현 및 정제 (Feature Representation and Cleaning)

- 벡터 공간 표현 (vector-space representation)
  - 역문서빈도 (inverse document frequency: IDF)
  - n: 문서의 수
  - i: 키워드 인덱스
  - n_i: 키워드 i가 발견된 문서 수

$$id_i = \log(n/n_i)$$

  - 특정 문서의 tf-idf 표현
    - $f(x_i)$: 키워드 i가 문서 내 등장한 횟수

$$h(x_i) = f(x_i)id_i$$

4.3.3 Collecting User Likes and Dislikes

- 사용자의 아이템에 대한 선호/비서도 정보 수집
  - 평점
  - 암시적 피드백
  - 텍스트 의견
  - 사례
- 수집된 데이터로부터 클래스 레이블 (class label) 생성
  - 학습에 활용

4.3.4 피쳐 선택과 가중치 설정 (Supervised Feature Selection and Weighting)

- 피쳐가 매우 많을 경우
  - 노이즈에 가까워 분류/추천에 도움이 안되는 피쳐 제거
  - 예) 키워드:수십만 ~ (반면에 평점이 매겨진 엔트리 수는 부족)
  - 정보성이 강한 상위 피쳐만 남김
- 비지도 방법 (Unsupervised methods)
  - 불용어 제거, 역문서 빈도 등
- 지도 방법 (Supervised methods)
  - 지니 계수, 엔트로피, Chi-square 등

4.3.4.1지니계수 (Gini Index)

- 세팅
  - t: 가능한 평점 수 (예: 1 선호, 2 비선호, 3 중립)
  - w: 특정 키워드
  - $p_1(w)$: 키워드 w가 포함된 아이템 중에서 평점이 1 (선호)인 아이템의 비율
- 지니 계수

$$\mathrm{Gini}(w) = 1 - \sum_{i=1}^{t} p_i(w)^2$$

  - 값의 범위: (0, 1-1/t)
  - 작을 수록 변별력 (discriminative power)가 강함
  - 예) $p_1(w)=1$, $p_2(w)=0$, $p_3(w)=0$
    - 지니계수: 0 (이 키워드가 등장한다면 평점이 1일게 거의 확실)
  - 예) $p_1(w)=1/3$, $p_2(w)=1/3$, $p_3(w)=1/3$
    - 지니계수: 2/3 (정보성이 없음)

4.3.4.2엔트로피 (entropy)

- 세팅
  - t: 가능한 평점 수 (예: 1 선호, 2 비선호, 3 중립)
  - w: 특정 키워드
  - $p_1(w)$: 키워드 w가 포함된 아이템 중에서 평점이 1 (선호)인 아이템의 비율
- 엔트로피

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

  - 값의 범위: (0, log$t$)
  - 작을 수록 변별력 (discriminative power)가 강함
  - 예) $p_1(w)=1$, $p_2(w)=0$, $p_3(w)=0$
    - 엔트로피: 0 (이 키워드가 등장한다면 평점이 1일게 거의 확실)
  - 예) $p_1(w)=1/3$, $p_2(w)=1/3$, $p_3(w)=1/3$
    - 엔트로피: log(3) (정보성이 없음)

4.3.4.3 카이제곱 통계치 (Chi-square Statistic)

- 세팅
  - 분할표(Contingency table)
    - 키워드와 클래스간 동시발생 (co-occurrence) 계산
  - 예상치 (expected numbers):l
    - 10%: 사용자가 상품을 구매할 가능성
    - 20%: 상품 설명에 키워드 w가 등장할 가능성
    - 동시 발생간 독립이라면…

	Term occurs in description	Term does not occur
User bought item	$1000 * 0.1 * 0.2 = 20$	$1000 * 0.1 * 0.8 = 80$
User did not buy item	$1000 * 0.9 * 0.2 = 180$	$1000 * 0.9 * 0.8 = 720$

  - 관측치 (observed numbers)

	Term occurs in description	Term does not occur
User bought item	$O_1 = 60$	$O_2 = 40$
User did not buy item	$O_3 = 140$	$O_4 = 760$

4.3.4.3 카이제곱 통계치 (Chi-square Statistic)

- 카이-스퀘어 통계치 계산

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

$$\begin{aligned}\chi^2 &= \frac{(60-20)^2}{20} + \frac{(40-80)^2}{80} + \frac{(140-180)^2}{180} + \frac{(760-720)^2}{720} \\ &= 80 + 20 + 8.89 + 2.22 \\ &= 111.11\end{aligned}$$

	Term occurs in description	Term does not occur
User bought item	$1000 * 0.1 * 0.2 = 20$	$1000 * 0.1 * 0.8 = 80$
User did not buy item	$1000 * 0.9 * 0.2 = 180$	$1000 * 0.9 * 0.8 = 720$

	Term occurs in description	Term does not occur
User bought item	$O_1 = 60$	$O_2 = 40$
User did not buy item	$O_3 = 140$	$O_4 = 760$

$$\chi^2 = \frac{(O_1 + O_2 + O_3 + O_4) \cdot (O_1 O_4 - O_2 O_3)^2}{(O_1 + O_2) \cdot (O_3 + O_4) \cdot (O_1 + O_3) \cdot (O_2 + O_4)}$$

4.3.4.3 카이제곱 통계치 (Chi-square Statistic)

- 실전
  - 지니 계수, 엔트로피, 카이제곱 통계치 등…
  - 각 피처마다 각 점수를 계산
  - 점수에 기반하여 피처를 정렬
    - 지니 계수, 엔트로피: 오름차순
    - 카이제곱: 내림차순
  - 상위 K개의 피처를 선택 (피처는 수십만~ 이상이기 때문에 선별 필요)
  - 모델 학습

4.3.4.5 가중치 설정 (Feature Weighting)

- 지니계수 등의 점수로 피처를 제외하지 않고
- 가중치로 삼아 모델 학습에 반영

4.4 사용자 프로파일 학습 (Learning User Profiles and Filtering)

- 세팅
  - 문서 (document): 상품을 설명하는 키워드의 나열
  - 클래스 레이블: 평점, 사용자 피드백으로부터 구성한 값
  - D_L: 레이블이 매겨진 문서 집합
  - 활성 사용자 (active user): 추천 시스템을 이용하는 사용자
  - 학습 모델
    - 활성 사용자 한명에 대한 문서 집합을 모아 학습
    - 적용 또한 해당 활성 사용자에 대해 실시
    - 모델이 바로 사용자 프로파일로 볼 수 있음
  - D_U: 레이블이 없는 테스트 문서 집합

4.4.1 최근접 이웃 분류 (Nearest Neighbor Classification)

- 스킵

4.4.3 베이즈 분류 모델

- 나이브 베이즈 분류 모델
  - 분류 (Classification) 문제의 베이스라인 모형 (심플함과 해석력이 장점)
  - 3장의 나이브 베이즈 협업 필터링의 기본이 됨
- 세팅
  - D_L: 레이블된 문서 집합
    - 레이블은 1 (선호), -1 (비선호) 라고 가정
    - $\bar{X} = (x_1, \ldots, x_d)$: 문서는 d개의 키워드로 표현
      - 키워드가 문서 내 존재 하는지 여부만 따짐 (베르누이 (Bernoulli) 모델)
      - 등장했으면 1, 그렇지 않으면 0
  - D_U: 레이블 없는 테스트 집합
- 목표

$$P(\text{Active user likes } \overline{X} | x_1 \ldots x_d)$$

4.4.3 베이즈 분류 모델

- 목표 $P(c(\overline{X}) = 1|x_1 \dots x_d)$ and $P(c(\overline{X}) = -1|x_1 \dots x_d)$

$$P(c(\overline{X}) = 1|x_1 \dots x_d) = \frac{P(c(\overline{X}) = 1) \cdot P(x_1 \dots x_d|c(\overline{X}) = 1)}{P(x_1 \dots x_d)}$$

$$\propto P(c(\overline{X}) = 1) \cdot P(x_1 \dots x_d|c(\overline{X}) = 1)$$

$$= P(c(\overline{X}) = 1) \cdot \prod_{i=1}^{d} P(x_i|c(\overline{X}) = 1) \quad \text{[Naive Assumption]}$$

  - Naive Assumption
    - 조건부 독립: 사용자가 선호 (비선호 문서를 제외)한 문서들만 골라놓고 보면, 각 키워드의 발생 확률은 서로 독립이다.
    - 즉, 단순히 선호 문서들 중 각 키워드가 있는/없는 문서 비율만 세면 됨

4.4.3 베이즈 분류 모델

- 계산 $$P(c(\overline{X}) = 1|x_1 \dots x_d) \propto P(c(\overline{X}) = 1) \cdot \prod_{i=1}^{d} P(x_i|c(\overline{X}) = 1)$$

$$P(c(\overline{X}) = -1|x_1 \dots x_d) \propto P(c(\overline{X}) = -1) \cdot \prod_{i=1}^{d} P(x_i|c(\overline{X}) = -1)$$

  - Prior: $P(c(\overline{X}) = 1)$
    - D_L 문서집합 중 `선호` 레이블이 된 문서의 비율
  - 조건부 확률: $P(x_i|c(\overline{X}) = 1)$
    - `선호` 레이블이 된 문서 중 키워드 i가 있는/없는 문서의 비율
- 라플라시안 스무딩 (Laplacian smoothing)

$$P(c(\overline{X}) = 1) = \frac{|\mathcal{D}_L^+| + \alpha}{|\mathcal{D}_L| + 2 \cdot \alpha} \qquad P(x_i|c(\overline{X}) = 1) = \frac{q^+(x_i) + \beta}{|\mathcal{D}_L^+| + 2 \cdot \beta}$$

  - 과적합 및 분자, 분모가 0이 될 수 있는 것을 방지

4.4.3 베이즈 분류 모델 (예)

Table 4.1: Illustration of the Bayes method for a content-based system

Keyword ⇒ / Song-Id ⇓	Drums	Guitar	Beat	Classical	Symphony	Orchestra	Like or Dislike
1	1	1	1	0	0	0	Dislike
2	1	1	0	0	0	1	Dislike
3	0	1	1	0	0	0	Dislike
4	0	0	0	1	1	1	Like
5	0	1	0	1	0	1	Like
6	0	0	0	1	1	0	Like
Test-1	0	0	0	1	0	0	?
Test-2	1	0	1	0	0	0	?

$P(x_1=0|\text{Like}) = \frac{3}{3}$

$P(x_2=0|\text{Like}) = \frac{2}{3}$

$P(x_3=0|\text{Like}) = \frac{3}{3}$

$P(x_4=1|\text{Like}) = \frac{3}{3}$

$P(x_5=0|\text{Like}) = \frac{1}{3}$

$P(x_6=0|\text{Like}) = \frac{1}{3}$

$$P(\text{Like}|\textit{Test-1}) \propto 0.5 \prod_{i=1}^{6} P(x_i|\text{Like})$$

주의: 해당 표는 한명의 활성 사용자에 대한 데이터임

$$= (0.5) \cdot \frac{3}{4} \cdot \frac{2}{2} \cdot \frac{3}{4} \cdot \frac{3}{3} \cdot \frac{1}{4} \cdot \frac{1}{3}$$

$$= \frac{3}{128}$$

$0.5 \times \frac{3}{3} \times \frac{2}{3} \times \frac{3}{3} \times \frac{3}{3} \times \frac{1}{3} \times \frac{1}{3} = \frac{1}{27}$

4.4.3 베이즈 분류 모델 (예)

Table 4.1: Illustration of the Bayes method for a content-based system

Keyword ⇒ / Song-Id ⇓	Drums	Guitar	Beat	Classical	Symphony	Orchestra	Like or Dislike
1	1	1	1	0	0	0	Dislike
2	1	1	0	0	0	1	Dislike
3	0	1	1	0	0	0	Dislike
4	0	0	0	1	1	1	Like
5	0	1	0	1	0	1	Like
6	0	0	0	1	1	0	Like
Test-1	0	0	0	1	0	0	?
Test-2	1	0	1	0	0	0	?

$P(x_1=0|\text{Dislike}) = \frac{1}{3}$

$P(x_2=0|\text{Dislike}) = \frac{0}{3}$

$P(x_3=0|\text{Dislike}) = \frac{1}{3}$

$P(x_4=1|\text{Dislike}) = \frac{0}{3}$

$P(x_5=0|\text{Dislike}) = \frac{3}{3}$

$P(x_6=0|\text{Dislike}) = \frac{2}{3}$

$$P(\text{Dislike}|\textit{Test-1}) \propto 0.5 \prod_{i=1}^{6} P(x_i|\text{Dislike})$$

$$= (0.5) \cdot \frac{1}{4} \cdot \frac{0}{2} \cdot \frac{1}{4} \cdot \frac{0}{3} \cdot \frac{3}{4} \cdot \frac{2}{3} = 0.5 \times \frac{1}{3} \times \frac{0}{3} \times \frac{1}{3} \times \frac{0}{3} \times \frac{3}{3} \times \frac{2}{3} = 0$$

$$= 0$$

4.4.4 규칙기반분류모델 (Rule-based Classifiers)

- 콘텐츠기반 모델로서의 규칙기반 모델
  - 연관규칙 (Association Rule): X ⇒ Y
    - Antecedent: 키워드가 상품 설명에 존재함
    - Consequent: 상품에 대한 평점
  - 예
    - {Classical, Symphony}
      - Support: 2 건 / 6행 = 33%
    - {Classical, Symphony, Like}
      - Support: 2건 / 6행 = 33%
    - {Classical, Symphony} ⇒Like
      - Confidence: 33%/33% = 100%
- 학습 단계
  - D_L에 대해 연관 규칙 생성
- 테스트 아이템에 대한 평점 예측
  - 해당 아이템이 antecedent 키워드를 포함하고 있는 규칙들을 검색
  - 규칙들의 consequent들의 평균으로 평점을 결정

Table 4.1: Illustration of the Bayes method for a content-based system

Keyword ⇒ / Song-Id ⇓	Drums	Guitar	Beat	Classical	Symphony	Orchestra	Like or Dislike
1	1	1	1	0	0	0	Dislike
2	1	1	0	0	0	1	Dislike
3	0	1	1	0	0	0	Dislike
4	0	0	0	1	1	1	Like
5	0	1	0	1	0	1	Like
6	0	0	0	1	1	0	Like
Test-1	0	0	0	1	0	0	?
Test-2	1	0	1	0	0	0	?

주의: 해당 표는 한명의 활성 사용자에 대한 데이터임

4.4.4 규칙기반분류모델 (Rule-based Classifiers)

- 예)
  - Support level: 33%, confidence level 75% 이상으로 한정시

Rule 1: {Classical} ⇒ Like (50%, 100%)

Rule 2: {Symphony} ⇒ Like (33%, 100%)

Rule 3: {Classical, Symphony} ⇒ Like (33%, 100%)

Rule 4: {Drums, Guitar} ⇒ Dislike (33%, 100%)

Rule 5: {Drums} ⇒ Dislike (33%, 100%)

Rule 6: {Beat} ⇒ Dislike (33%, 100%)

Rule 7: {Guitar} ⇒ Dislike (50%, 75%)

Table 4.1: Illustration of the Bayes method for a content-based system

Keyword ⇒ Song-Id ⇓	Drums	Guitar	Beat	Classical	Symphony	Orchestra	Like or Dislike
1	1	1	1	0	0	0	Dislike
2	1	1	0	0	0	1	Dislike
3	0	1	1	0	0	0	Dislike
4	0	0	0	1	1	1	Like
5	0	1	0	1	0	1	Like
6	0	0	0	1	1	0	Like
Test-1	0	0	0	1	0	0	?
Test-2	1	0	1	0	0	0	?

- 예측시 테스트 아이템에 매칭되는 규칙 선택하는 방법의 예시
  - Rule 1: Test-1에 대해 선택됨 (Rule 1의 antecedent 키워드 전부가 Test-1의 키워드에 포함됨: {Classical} ⊂ {Classical})
  - Rule 5, 6:Test-2에 대해 선택됨 (Rule 5, 6의 antecedent 키워드 전부가 Test-2의 키워드에 포함됨: {Drums} ⊂ {Drums, Beat}

4.4.5 회귀 기반 모델 (Regression-Based Models)

- 세팅
  - D_L: nxd 행렬 (n documents, d keyword features)
  - $\bar{y}$: n 차원 열 벡터 (n-dimensional column vector)
    - n 개의 문서(상품)에 대한 평점
  - $\overline{W}$: d 차원 행 벡터 (d-dimensional row vector)
    - 각 키워드에 대한 선형 결합 계수
- 선형 회귀 (linear regression): $\overline{y} \approx D_L \overline{W}^T$
  - Prediction error $(D_L \overline{W}^T - \overline{y})$
  - Regularization term $\lambda ||\overline{W}||^2$

4.4.5 회귀 기반 모델 (Regression-Based Models)

- 최적화

$$\text{Minimize } O = ||D_L\overline{W}^T - \overline{y}||^2 + \lambda||\overline{W}||^2$$

- Closed form solution

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

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

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

4.5 Content-Based vs CF

- 콘텐츠 기반의 장점
  - 새로운 아이템의 추가에 대응 가능 (콜드 스타트 문제)
  - 상품의 피처에 대한 설명력 제공 가능
  - 기존 텍스트 분류기를 통해 구현 가능
- 콘텐츠 기반의 단점
  - Overspecialization: 사용자가 경험한 상품과 유사한 것만 발견하는 경향
    - 참신함 (Novelty), 의외성 (serendipity) 약점으로 이어짐
    - CF의 경우는 피어 그룹의 활용으로 일부 상쇄
  - 새로운 사용자의 추가에 대응 어려움 (콜드 스타트 문제)

4.6 Using Content-Based Models for CF

- 콘텐츠 기반 방법을 CF에 활용 가능
  - 예) 사용자이름과 키워드 연결하여 `새로운 키워드`를 생성

Terminator: John#Like, Alice#Dislike, Tom#Like
Aliens: John#Like, Peter#Dislike, Alice#Dislike, Sayani#Like
Gladiator: Jack#Like, Mary#Like, Alice#Like

- 사용자 프로파일을 키워드화 할 수 있는 경우
  - 이를 입력에 포함시켜 하나의 글로벌 분류만을 학습/활용
    - (모델 사용자 마다 학습 모델을 만들지 말고)
    - 최근의 흐름

4.7 요약

- 콘텐츠 기반 추천 시스템
  - 상품 설명을 콘텐츠 속성화하고
  - 사용자 평점과 결합하여
  - 사용자 프로필 생성 (모델 학습)
  - 이를 활용하여 테스트 상품에 대한 평점 예측
- 연습해볼 것
  - 피쳐 선택 계산 방법 (지니계수, 엔트로피, 카이제곱)
  - 연관규칙 계산방법
  - 나이브베이즈 계산방법