제4장 콘텐츠 기반 추천 시스템

4.1 개요

콘텐츠 기반 추천 시스템은 **설명 가능한 속성 집합**을 이용해 아이템을 설명할 수 있는 시나리오를 활용하도록 설계되었습니다. 이 방법은 특히 새로운 아이템이거나 해당 아이템에 대한 평가가 거의 없는 경우 유용합니다.

콘텐츠 기반 시스템은 사용자가 좋아하는 아이템의 속성 유사도를 기반으로 **사용자가 과거에 좋아했던 것과 유사한 아이템**을 추천합니다. 가장 기본적으로 다음과 같은 데이터를 사용합니다.

1. 콘텐츠 중심 속성의 다양한 아이템에 대한 설명
2. 다양한 아이템에 대한 사용자 피드백에서 생성된 사용자 프로파일

일반적으로 콘텐츠 기반 추천 시스템에서는 다른 사용자의 평점은 아무런 역할을 하지 않습니다. 이는 컨텍스트에 따라 장점이기도 하고 단점이기도 합니다. 정보가 거의 없는 콜드 스타트의 경우 이용 가능하다는 대표적인 장점이 있지만 추천의 다양성과 참신성이 떨어질 수 있다는 단점도 공존합니다.

콘텐츠 기반 시스템은 주로 많은 양의 속성 정보를 즉시 사용할 수 있는 시나리오에서 사용하는데 대부분의 경우 아이템 설명에서 추출된 키워드(텍스트 속성)입니다. 웹 페이지 추천이나 이커머스 환경에서 매우 유리합니다.

또한, 콘텐츠 기반 시스템은 지식 기반 시스템과 밀접한 관련이 있습니다. 지식 기반 추천 시스템은 추천으 만들기 위해 아이템의 콘테느 속성을 사용합니다. 두 추천 시스템의 차이점은 지식 기반 시스템은 사용자가 추천 관련 요구 사항을 명시적으로 입력하도록 하지만 콘텐츠 기반 시스템은 일반적으로 과거 평가를 학습하는 접근 방식을 사용합니다.

4.2 콘텐츠 기반 시스템의 기본 구성 요소

콘텐츠 기반 시스템은 아이템에 대한 설명 및 지식을 사용하기 때문에 여러 유형의 **비정형 데이터**를 **표준화된 설명**으로 변환해야 합니다. 대부분 keyword로 변환하는 것이 선호되며 많은 경우 텍스트 중심으로 작동합니다.

콘텐츠 기반 시스템의 주요 구성 요소는 전처리, 학습 및 예측 부분을 포함합니다.

전처리 피처 추출: 콘텐츠 기반 시스템은 웹 페이지, 제품 설명, 뉴스 음악 등과 같은 다양한 영역에서 사용합니다. 다양한 source에서 feature를 추출해 keyword 기반 벡터 공간(verctor space)으로 변환합니다. 유익한 feature를 적절하게 추출하는 것은 콘텐츠 기반 추천 시스템의 효과적인 기능을 위해 필수적입니다.

사용자 프로필의 콘텐츠 기반 학습: 사용자 관심을 아이템 속성과 관련시키기 때문에 사용자 프로필로 명명합니다. 콘텐츠 기반 모델은 사용자에 따라 다른 추천을 제공하고 사용자의 평점과 같은 명시적 피드백 또는 사용자 활동과 같은 암시적 피드백을 활용합니다.

필터링 및 추천: 이전 단계에서 학습된 모델을 사용해 특정 사용자의 아이템에 대한 추천을 만듭니다. 해당 단계는 온라인으로 이뤄지기 때문에 효율성이 중시됩니다.

4.3 전처리 및 피처 추출

모든 콘텐츠 기반 모델의 첫 번째 단계는 아이템 사이 차이를 분간할 수 있는 feature를 추출하는 것입니다. Feature는 사용자의 관심 분야를 예측하는 데 큰 도움을 제공합니다.

4.3.1 피처 추출

피처 추철 단계에서는 다양한 아이템에 대한 설명을 추출합니다. 가장 일반적인 방법은 기본 데이터에서 키워드를 추출하는 것입니다. 분류 과정에서 쉽게 사용하기 위해 다양한 필드에 대한 가중치를 적절하게 부여할 필요가 있으며 피처 가중치 설정은 피처 설정과 밀접한 관련이 있습니다.

4.3.1.1~3 애플리케이션별 예시

상품 추천 - 영화에 대한 개인화된 추천 등, 웹 페이지 추천, 음악 추천

4.3.2 피처 표현 및 정제

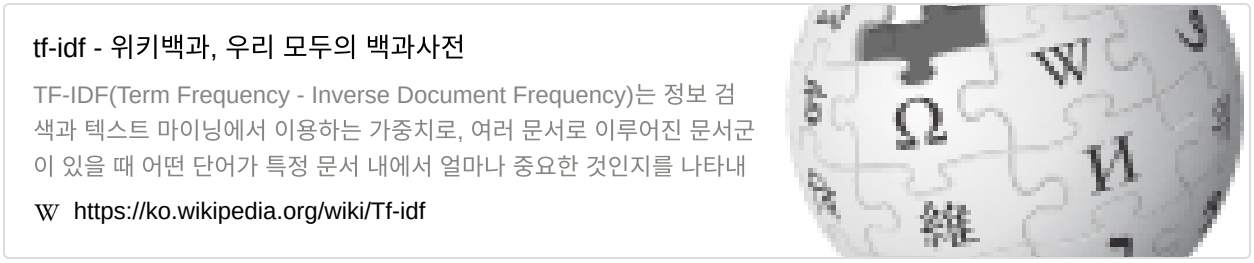
피처 표현 및 정제 단계는 구조화되지 않은 형식의 표현을 위해 사용할 때 더욱 중요합니다. 피처 추출 단계에서 keyword 모음을 결정할 수 있습니다. 그러나 이러한 표현은 처리하기에 적합한 형식으로 정리하고 표현해야 합니다. 다음은 정제 과정에서 거치는 몇 가지 단계입니다.

1. **불용어 제거:** ‘a’, ‘an’와 ‘the’ 같은 단어는 대부분 문장에서 빈도 높게 찾아볼 수 있습니다. 통상 관사, 전치사, 접속사 및 대명사는 불용어로 취급하며 대부분의 경우 표준화된 불용어 목록은 영어 뿐만 아니라 다양한 언어로 제공됩니다.
2. **형태소 분석:** 동일한 단어의 유사어를 통합합니다. 예를 들어 단수 또는 복수형의 단어 또는 동일한 단어의 다른 시제를 통합합니다. 공통 어근을 추출하기도 하며 이 또한 많은 툴이 존재합니다.
3. **구문 추출:** 문서에서 함께 자주 발생하는 단어를 검새합니다. 동일한 어휘라도 구성 단어에 따라 다른 것을 의미할 수 있기 때문에 자동화된 방법을 사용하거나 수동으로 정의된 사전을 구문 추출에 사용할 수 있습니다.

위와 같은 단계를 실행하고 키워드를 벡터 공간 표현으로 변환합니다. Vector space 표현에서 문서를 단어 frequency와 함께 ‘bags of words’로 표현합니다. 빈도가 낮은 단어를 완전히 제거

하는 방식 대신 의미 없는 것으로 치부해 부드러운 방식으로 수행한다는 점을 제외하고는 불용어의 원칙과 유사합니다.

단어를 가치가 없는 것으로 처리하려면 **역문서빈도**(idf: inverse document frequency)의 개념을 사용해야 합니다. 활용된 가장 일반적인 모델은 **tf-idf 모델**이라고 하며, 여기서 tf는 term frequency, idf는 inverse document frequency를 의미합니다.



4.3.3 사용자가 좋아하는 것과 싫어하는 것 수집

콘텐츠 정보 이외에도 추천 과정에서 사용자의 좋아하는 것과 싫어하는 것에 대한 데이터를 수집해야 합니다. 데이터 수집은 오프라인 단계에서 수행하지만 추천은 특정 사용자가 시스템과 상호작용하는 온라인 단계에서 결정됩니다. 수집하는 데이터는 다음 형식 중 하나를 택할 수 있습니다.

평점: user는 item에 대한 선호도를 나타내는 평점을 지정합니다.

암시적 피드백: item 구매 또는 검색과 같은 사용자 행동을 나타냅니다.

텍스트 의견: 의견 마이닝 및 감정 분석 분야와 관련 있습니다.

사례: user는 관심 있는 아이템의 예를 명시할 수 있습니다.

4.3.4 지도 피처 선택과 가중치 설정

피처 선택 및 가중치 설정의 목표는 가장 유익한 단어만 벡터 공간에 남도록 하는 것입니다. 여러 영역에서 수행한 실험 결과, 추출된 단어의 수가 50~300 사이에 있어야 한다고 제안합니다. Noise가 많은 단어는 과적합되는 결과를 만드는 경우가 많고, 학습에 ㅏ용할 수 있는 문서의 수가 적으면 모델이 과적합되는 경향도 커집니다.

문서 표현에 피처 정보를 통합할 때 두 가지 뚜렷한 측면이 있습니다.

하나는 **피처 선택**으로, 단어 제거에 해당합니다.

두 번째는 단어에 더 큰 중요성을 부여하는 **피처 가중치 설정**입니다.

하지만 이는 사용자의 feedback이 중요하지 않은 unsupervised feature selection 및 weight setting이기 때문에 사용자 평점을 고려한 feature 선택을 위한 supervised method을 학습합니다.

4.4 사용자 프로필 학습 및 필터링

사용자 프로파일 학습은 분류 및 회귀 모델링 문제와 밀접하게 관련되어 있습니다. 평점을 '좋아요' 또는 '싫어요'처럼 개별 값으로 처리하면 텍스트 분류와 유사하고 반면, 평점을 수치 값으로 취급하면 문제는 회귀 모델링과 유사합니다.

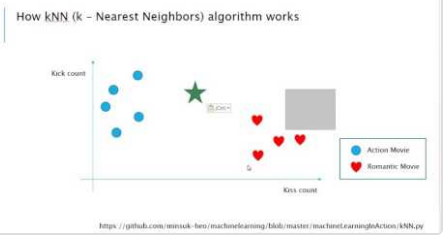
4.4.1 최근접 이웃 분류

최근접 이웃 분류(Nearest Neighbor Classifier)는 가장 단순한 분류 기술 중 하나이며 비교적 직접적인 방법으로 구현할 수 있습니다.

[머신러닝] kNN(k-Nearest Neighbors) 최근접 이웃 알고리즘

kNN 첫번째 강의입니다. kNN 알고리즘을 이해하기 위한 강의입니다.액션 영화와 로맨틱 영화라는 기존 데이터를 통해 새로 나온 영화가 액션 영화인지 로맨틱 영화인지 판별하는 머신 러닝 알고리즘 예제를 다룹니다.제

https://www.youtube.com/watch?v=CyuI2F_wJWw



4.4.2 사례 기반 추천 시스템

최근접 이웃 방법은 일반적으로 지식 기반 추천 시스템, 특히 사례 기반 추천 시스템과 연결됩니다. 사례 기반 추천 시스템은 사용자가 대화형으로 관심 있는 한 개의 관심 있는 사례를 선택하면 사용자가 관심 갖는 해당 사례의 가능한 아이템들을 받을 수 있습니다.

4.4.3 베이즈 분류 모델

*추후 자세히 기재 예정

4.4.4 규칙 기반 분류 모델

규칙 기반 분류 모델은 leave-one-out 메서드와 연관 메서드를 비롯한 다양한 방식으로 디자인할 수 있습니다.

콘텐츠 기반 시스템의 규칙 기반 분류 모델은 협업 필터링의 규칙 기반 분류와 유사합니다. 협업 필터링의 아이템-아이템 규칙에서 선행 과목과 결과 규칙은 아이템의 평점에 해당합니다. 주요 차이점은 협업 필터링에서 규칙의 선행은 다양한 아이템의 평가에 해당하는 반면, 콘텐츠 기반 방법에서는 규칙의 선행 아이템이 아이템 설명에서 특정 키워드의 존재와 일치한다는 것입니다.

4.4.5 회귀 기반 모델

회귀 기반 모델은 binary 평가, 구간 기반 평가 또는 수치 평가와 같은 다양한 유형의 평가에 사용할 수 있다는 장점이 있습니다. 선형 모델, 로지스틱 회귀 모델 및 순서형 프로빗 모형과 같은 대규모 회귀 모델을 사용해 다양한 유형의 평점을 모델링할 수 있습니다.

4.4.6 기타 학습 모델 및 비교 개요

콘텐츠 기반 필터링의 문제는 분류 및 회귀 모델링을 직접 적용하는 것이므로 문헌에서 많은 다른 기술이 사용될 수 있다는 것입니다.

4.4.7 콘텐츠 기반 시스템에 관한 설명

콘텐츠 기반 시스템은 콘텐츠 피처를 기반으로 모델을 추출하기 때문에 추천 프로세스를 통찰력 있게 해석하는 경우가 많습니다. 협업 필터링 추천 시스템에서는 세부적인 설명이 부족한 경우가 많으며, 추천은 아이템의 상세한 특징보다는 유사한 아이템을 통해서 설명할 수 있습니다.

그러나 통찰력의 성격과 정도는 사용한 특정 모델에 매우 민감합니다.

4.5 콘텐츠 기반 대 협업 필터링 추천

다음은 콘텐츠 기반 시스템의 장점입니다.

1. 모든 추천 시스템은 콜드 스타트 문제가 있지만 콘텐츠 기반 시스템은 새로운 사용자에 대해서만 콜드 스타트 문제가 발생합니다.
2. 아이템의 피처 측면에서 설명이 가능합니다.
3. 일반적으로 상용 텍스트 분류 모델과 함께 사용할 수 있습니다.

다음은 콘텐츠 기반 시스템의 단점입니다.

1. 과적합 문제가 있습니다.
2. 신규 사용자에 대한 콜드 스타트 문제가 치명적입니다.

4.6 협업 피터링 시스템을 위한 기반 모델 사용

협업 필터링 모델과 콘텐츠 기반 방법은 흥미롭게 연결되어 있습니다. 콘텐츠 기반 방법은 협업 필터링에서 직접 사용할 수 있습니다. 아이템의 콘텐츠 설명은 설명 키워드를 참조하지만 사용자의 평점을 활용해 콘텐츠 기반 설명을 정의하는 시나리오를 계획할 수 있습니다.

4.7 요약

아이템 설명의 콘텐츠 속성은 사용자 평점과 결합돼 사용자 프로필을 작성합니다.

최근접 이웃 분류 모델, 규칙 기반 방법, 베이즈 방법 및 전형 모델과 같은 많은 분류 및 회귀 모델을 해당 시스템에 사용하고 그중에서도 특히 베이즈 방법은 다양한 유형의 콘텐츠를 처리할 수 있기 때문에 다양한 시나리오에서 큰 성공을 거뒀습니다.

콘텐츠 기반 시스템은 신규 사용자와 관련해 콜드 스타트 문제를 해결할 수 없지만 **새 아이템과 관련된 콜드 스타트 문제**를 처리할 수 있다는 장점이 있습니다.

콘텐츠 기반 시스템은 사용자가 이전에 평가한 아이템의 내용을 기반으로 하므로 콘텐츠 기반 시스템에서 의외성을 주는 아이템이 추천될 확률은 상대적으로 낮습니다.