**추천시스템**

1. 추천이란?

추천은 예측이다. 추천은 상대방이 선호할만한 결과를 제안하는 것이다. 사용자에게 정보를 제공했을 때 과연 어떤 반응을 보일 것인가를 예측하는 것 또한 추천이 할 일이다. 긍정적인 반응을 끌어올 만한 것이 추천의 대상이다.

추천은 결과를 얻고자 하는 대상을 분석하는 일부터 시작한다. 그것은 대상에게 주고자 하는 것을 분석하는 일이 될 수도 있고, 그 대상을 분석하는 일이 될 수도 있다. (전통주를 추천한다 🡪 전통주를 소비하는 소비자 or 전통주를 분석한다.)

추천시스템은 이러한 추천을 데이터적이고, 컴퓨터적으로 옮겨온 것이다. 또한, 데이터의 양이 아주 많아지고 유의미한 데이터를 분석하는 기법이 발달함에 따라 추천시스템을 구축할 수 있게 되었다.

하지만 추천은 데이터적으로 완전하게 맞는 정답이 나오기는 힘들다. 왜냐하면, 결과로 보여준 것들의 대부분은 사용자가 경험하지 못한 것들에 대한 것이기 때문이다. 이는 결과를 받은 후에, 평가로 확인하여야 비로소 정확한지 아닌지를 따질 수 있는 것이다.

추천이라는 분야는 이제 비로소 머신러닝 분야로 속하게 되었다. 데이터와 관련된 분석 및 예측을 주로 이용하기 때문이다. 현재의 추천 시스템은 예측과 피드백 데이터를 바탕으로 모델 성능을 평가한다. **추천시스템에서 더 나은 평가를 받기 위해서는 적절한 피드백을 받아야 하며, 이 피드백이 중요하다고 생각한다.**

2. 추천시스템의 목적

관련성: 사용자에게 적합한 아이템을 추천

참신성: 사용자가 전에 본 적 없던 새로운 아이템을 추천

우연성: 사용자에게 다소 의외이거나 놀라움을 주는 아이템을 추천

3. 추천 시스템의 원리

1. **20대 여성들은 분위기 좋은 카페를 선호한다.**

* **20대 여성들을 대상으로 설문조사를 하여 나온 데이터 분석 결과라고 볼 수 있습니다.**
* **이때, 어떤 사용자가 20대 여성일 때, 카페를 찾고자 한다면 시스템은 평범한 카페보다 상대적으로 분위기 좋은 근처의 카페를 추천할 수 있습니다. 1번을 근거로 말입니다.**

2. **20대 남성들은 액션이 많고 화려한 볼거리의 영화를 많이 보더라**

* **20대 남성이라는 타깃을 분석한 결과라고 생각하고,**
* **어떤 사용자가 20대 남성이면 위를 근거로 액션 영화를 추천할 가능성이 높다는 것이 바로 추천의 원리입니다.**

위와 같은 예시들도 다 추천이다. 하지만 위의 예시들은 문제점이 많다. 인기, 대중성에 기반하여 일반화된 추천을 제공한다는 점이 첫번째이다. 또한 개개인의 성향을 모르고, 큰 타깃 범위만으로 추천을 하는 것이다. 추천 결과에 대한 만족도가 떨어질 수 밖에 없는 이유이다. **(but, 대중화, 상업화가 필요한 아이템, 즉 소비자가 모르는 아이템에 대해서는 큰 타깃 범위로 추천해주는 것이 맞다고 본다. Like 전통주)**

위의 예시와 같은 추천이 이루어지다가, 더 다양한 데이터 수집이 가능해지고, 개인의 데이터가 아주 많아지게 된다. 또한, 소비자의 욕구들이 아주 다양해졌다. 이것에 대응하기 위해서 맞춤형 정보를 제공하고자 한다. 기존의 추천 시스템보다 더 개인화되고 정확한 추천을 하고자하는 것이다. 그래서 추천 알고리즘이 나오게 된다.

4. 추천 알고리즘의 변화

기존의 단순한 알고리즘에서 다양한 대응을 할 수 있는 시스템의 구축을 위해 알고리즘이 더 복잡해진다.

어제의 데이터를 분석하여 내일의 사용자의 행동을 예측하는 머신 러닝 알고리즘으로 가게 되는 것이다. 이것은 다양한 데이터 소스를 활용하여 소비자의 취향과 원하는 것을 추론하고자 한다. 사용자와 아이템 간의 상호작용을 점수로서 계산하고, 점수가 높은 것을 추천하는 것이다.

대표적인 예가, 컨텐츠 기반 추천시스템, 협업 필터링 추천시스템이다.

5. 추천 시스템의 대표적인 모델

1. Prediction Version

- Matrix Completion Problem(행렬에 결측치가 있는 문제)

- 사용자와 아이템 간의 평가 점수를 예측하는 것

- M명의 user, n개의 item이 있을 떄, M\*N MATRIX를 계산하는 방식이다.

2. Ranking Version

- Top-K Recommendation Problem

- 특정 유저에게 k개의 적합한 아이템을 추천해주는 방식이다.

6. 컨텐츠 기반 추천 시스템

간단하게 말하면, 특정 아이템에 기초하여 비슷한 아이템을 추천해준다는 것이다.

-아이템끼리의 유사도를 측정한다.

-아이템의 metadata를 사용한다. Metadata is **data** that provides information about other **data**

Ex) 장르, 감독, 묘사, 배우 etc

-만약 사용자가 특정 아이템을 선호한다고 하면, 해당 아이템과 유사한 아이템을 추천하는 방식이다.

특징은 사람들 간을 비교하지 않고, 아이템 특성만을 비교한다는 점이다. 사용자가 적은 초반의 추천 시스템의 경우 대부분 Content-based recommender를 사용한다.

6-1. 장단점

장점

- 다른 사용자의 영향을 받지 않는다.

- 새로운 아이템에 대해서도 추천이 가능하다.

- 추천을 설명하기 쉽다.

단점

- 새로운 유저에게 추천이 불가능하다.

- 소리, 영상, 이미지 등의 콘텐츠로부터 추천을 위한 metadata, 특징을 추출하기가 어렵다.

6-2. 추천의 순서

1. 아이템의 분류를 통해 콘텐츠를 분석한다.

2. 각각의 유저의 선호를 대표하는 유저의 profile을 학습한다.

3. 아이템의 분류 & 사용자 선호도에 따라 각각의 유저를 위한 추천 리스트를 생성한다.

6-3. ML Techniques

사용되는 머신러닝 테크닉은 아래와 같다.

KNN Classification(nearest neighbor)

Linear Classification

🡪 **이둘은 언제 사용되는 거?**

6-4. 컨텐츠 묘사

구조화된 컨텐츠

구조화되지 않은 컨텐츠

- Ex) keyword

- BOW방식 또는 TF/IDF방식을 사용한다. 단어의 중요성을 검토하여 사용한다.

6-5. 아이템 유사도

Cosine Similarity

그 외 다른 적합한 유사도

7. 협업 필터링 추천시스템

사용자 그룹이 형성되어 있고, 그들 간의 평가 점수와 선호도를 고려하여, 사용자의 예측 점수와 선호도가 결정된다. 한마디로, 사용자와 비슷한 다른 사용자를 찾아서 그 사용자는 어떤 평가를 했는지를 살펴보는 방식이다.

사용자가 어느 정도 형성이 되어 있고, 데이터가 존재할 때 사용 가능한 추천 알고리즘 방식이다.

7-1. 장단점

장점

- 어떠한 아이템에 대해서도 추천이 가능하다. (아이템의 특성에 의존하지 않는다.)

단점

- 평가되지 않은 아이템에 대해 추천을 하지 않는다. (new – item problem)

- 보통 가장 인기 있는 아이템을 추천한다.

- 비슷한 유저 군이 존재하는 사용자 그룹이 어느 정도 숫자 이상 필요하다. (cold start problem for new users)

7-2. 작동 방식

INPUT

- 주어진 유저-아이템 간의 평가 점수 matrix

OUTPUT

- 특정 아이템에 대한 선호도 예측 점수

- TOP-N 추천 리스트

7-3. 두가지 방식의 CF

7-3-1. USER-BASED FILTERING

유저들의 평가 점수를 바탕으로, 예측하고자 하는 유저의 점수를 예측하는 방식

유저간의 유사성을 계산하고 예측한다.

7-3-1-1 장단점

장점

- 구현하기 쉽다.

- 문맥이 독립적이다.

- CONTENTS BASED에 비해 정확하다.

단점

- 희소성: 평가를 하는 사용자가 적다.

- 확장성: 사용자 그룹이 커질수록 비용이 비싸다.

- COLD START: 신규 사용자에게 추천이 힘들다.

7-3-2. ITEM BASED FILTERING

사용자 그룹을 사용하여, 관계를 판단하기에 CF에 포함된다.

아이템 사이의 유사성을 판단하여 예측을 한다. (KNN step을 건너뛴다.)

User-based보다 더 좋은 성능을 보인다(?)

요즘은 하이브리드를 많이 사용함.

실제 추천 시스템 사례 분석

1. 페이스북

* 1. 추천 시스템을 도입하게 된 문제점

페이스북에는 사용자와 연관 있는 사람들의 게시글인 뉴스피드라는 것이 있다. 초기에는 사용자당 읽어야 할 게시글이 적어서 문제가 없었다. 하지만 회원수가 증가함에 따라 뉴스피드의 양 역시 방대하게 증가했고 사용자는 관심 없는 글을 읽게 되고, 관심 있는 글은 너무 아래쪽에 있어서 못 읽게 되는 경우가 다반사였다. 그러므로 사용자에게 적절한 뉴스피드를 추천해주고자 이 추천 시스템을 도입하고자 했다.

* 1. 추천 시스템에 어떤 알고리즘을 사용햇나?

EDGE RANK ALGORITHM

사용자와 관련된 게시물을 먼저 보여주기 위해서 우선 순위를 매기는 알고리즘이다. 이는 사용자가 좋아할만한 뉴스피드를 우선적으로 노출시키는 추천 시스템이다.

세 가지 기준으로 점수를 부여

친밀도, 가중치, 시간을 기준으로 사용

친밀도 > 가중치> 시간 순서로 점수에 가중치를 부여

합산 점수의 순위가 높을수록 상위에 노출

### 친밀도 계산 방식

Affinity

**사용자와 게시글을 작성한 작성자 간의 관계를 살펴봅니다.**

**여기서 반응이란 좋아요, 댓글, 공유, 태그 등의 활동을 의미합니다.**

* **사용자 측면**
  + **작성자의 다른 게시물에 대한 반응**
  + **작성자 담벼락 게시물에 대한 반응**
* **작성자 측면**
  + **사용자의 게시물에 대한 반응**
  + **사용자 담벼락 게시물에 대한 반응**
* **사용자-작성자 간의 메신저 이용도**

### 가중치 계산 방식

Weight

**과거 사용자가 반응한 게시물과 현재 평가하고자 하는 게시물의 유사성 & 다른 사용자들의 게시물에 대한 반응을 살펴봅니다.**

**또한 아래의 우선순위로 가중치를 계산합니다.**

**동영상 > 그림/사진 > 텍스트**

**링크 > 직접 작성**

### 시간에 따른 계산 방식

Decay

**게시물이 작성된 시간, 게시물에 대한 마지막 반응 시간, 작성자와 관계를 맺은 시기 등을 기준으로 계산합니다.**

### 그 외의 추천 방식

* **스토리 범핑(Story Bumping) : 지나간 게시물 분석**
* **라스트 액터(Last Actor) : 최근 50개의 게시물 분석**
* **사용자가 직접 '먼저 보기' 버튼을 통해서 최상단에 노출**

Story Bumping & Last Actor  
  
출처: <https://davinci-ai.tistory.com/13?category=909551> [DAVINCI - AI]

2. 넷플릭스

넷플릭스에서 사용하는 알고리즘. (CINE-MATCH ALGORITHM)

컨텐츠기반 추천 + 협업 필터링 추천 🡪 하이브리드 추천 알고리즘

BAYESIAN CLASSIFIER(머신러닝 – 지도학습 - 분류 )

A영화를 본 사람이 B영화를 볼 확률 (큰 표본을 작은 표본으로 예측)

나이브 베이즈 분류는 베이즈 정리에 기반한 통계적 분류 기법이다. 가장 단순한 지도 학습 중 하나이다.

나이브 베이즈의 장단점

장점

- 간단하고, 빠르며, 정확한 모델이다.

- computation cost가 작다.(따라서 빠르다)

- 큰 데이터셋에 적합하다.

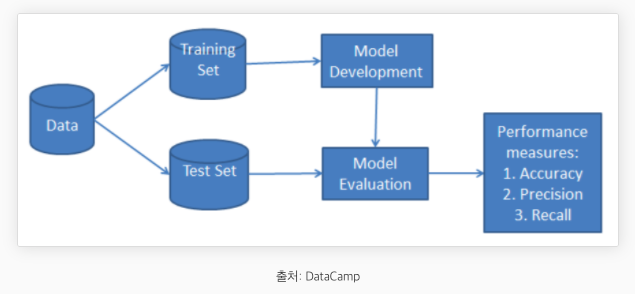
- 연속형보다 이산형 데이터에서 성능이 좋다.

- MULTIPLE CLASS 예측을 위해서도 사용할 수 있다.

단점

- feature간의 독립성이 있어야 한다. 하지만 실제 데이터에서 모든 feature가 독립인 경우는 드물다. 장점이 많지만 feature가 독립이어야 한다는 크리티컬한 단점이 있다.

나이브 베이즈는 feature끼리 서로 독립이라는 조건이 필요하다. (독립임을 어떻게 확인하지?)



해당 이미지는 지도학습을 했을 시의 흐름.

ASSOCIATION ANALYSIS: CO-OCCURRENCE (머신러닝 - 비지도 학습)

동시 발생 상관관계, A와 B를 동시에 좋다고 판단할 떄, A와 B의 관계를 알아보는 방법

지지도

어떤 규칙의 지지도가 10%라면 전체 트랜잭션 중 그 규칙을 따르고 있는 트랜잭션이 10%를 차지한다는 의미이다. 즉, 두 품목 A와 B의 지지도는 전체 거래 항목 중에서 항목 A와 B를 동시에 포함하는 거래의 비율을 말한다.

품목 A와 B를 포함하는 거래 수 / 전체 거래 수

SUPPORT = P (A교집합B)

신뢰도

신뢰도는 항목 A의 거래 중에서 항목 B가 포함된 거래의 비율을 말한다.

Confidence = P(B|A) = P(A교집합B)/P(A)

지지도(SUPPORT)/P(A)

향상도

A🡪 B의 연관 규칙에서 임의로 B가 구매되는 경우에 비해 A와의 관계가 고려되어 구매되는 경우의 비율이다. 즉, 연관 규칙이 오른쪽 항목을 예측하기 위한 능력이 얼마나 향상되었는가를 표현하는 값이다.

NEURAL NETWORK

소비자 선호 분야를 포함한 영화적 ATTRIBUTE들에 대해 가중치를 결정

### 사용 데이터

* **Input Data**
  + **기존의 영화 데이터 : 장르, 감독, 배우, 태그(줄거리, 분위기, 유머 코드, 등장인물 특성 등)**
  + **사용자의 개별 콘텐츠 소비 취향 : 찜 한 영화, 좋아요/싫어요**
  + **사용자의 시청 습관 : ex) 특정 콘텐츠를 중지한 시점, 빨리 넘기는 시점, 보지 않고 찜만 해 놓은 콘텐츠 등의 미세한 콘텐츠 소비 패턴**
* **중간 분석을 위한 생성 Data**
  + **사용자와 다른 사용자와의 영화 선택 유사성**
  + **사용자가 안 본 콘텐츠의 선호 예상 점수**
* **Output Data**
  + **추천 영화 리스트, 영화의 포스터**

### Tagger

Tag : 내용과 무관하게 Tag를 설명하기 위한 그림입니다.

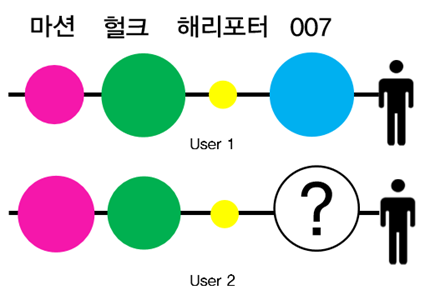
**넷플릭스는 콘텐츠를 약 5만여 가지의 태그 중 적합한 태그를 부여하여 정리한다고 합니다. 이렇게 나눈 콘텐츠를 개별 이용자의 이용 이력에 따라 맞춤형으로 더 정교한 추천을 해주는 것입니다.**

**이렇게 많은 종류의 태그를 콘텐츠에 부여하는 것은 직접 사람이 한다고 합니다. 넷플릭스는 50여 명의 정규직 태거가 근무하며, 일과에 넷플릭스 콘텐츠를 보고 직접 태그를 부여한 다고 합니다. 그래서 콘텐츠에 부여된 Tag들을 People Powered data라고 하기도 합니다.**

**이를 통해 우리가 알 수 있는 점은 정교한 추천을 위해서는 정교한 데이터가 필요하다는 것을 알 수 있고, 이는 아직 사람이 해야 하는 아주 힘든 작업이라는 점입니다.**

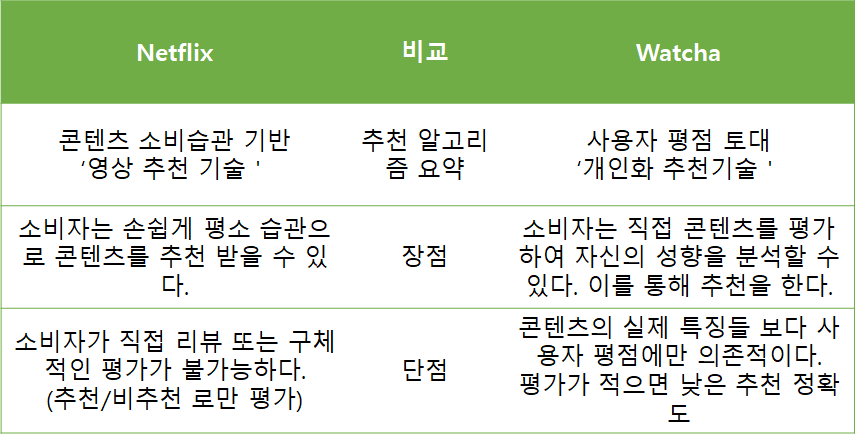
3. 왓차플레이

왓차플레이의 알고리즘 🡪 협업필터링 추천(선호하는 영화를 바탕으로, 사용자 간의 연관성을 파악한다)

Collaborative Filtering

**위의 그림에서 동그라미(선호도)의 크기는 유저가 결정(평점)합니다. 이를 통해서 유저 간의 경향성 (Covariance)을 파악합니다.**

### Netflix VS Watcha 알고리즘 비교

넷플릭스 왓챠 비교

4. 아마존의 추천시스템

자신들 만의 고유한 ITEM-TO-ITEM 협업 필터링 추천 시스템을 활용

- 제품들 간의 유사성을 계산하는 방식

- 특정 제품을 구매한 모든 고객에 대해서 각 고객이 구매한 다른 제품과 이 제품과의 유사성의 합을 계산한다.

- 이것들을 모든 고객들에게 적용시켜서 계산한다.

- 사용자의 검색 기록, 리뷰, 찜 목록, 장바구니, 구매한 물품 등의 사용자 정보를 전체 사용자와 비교하는 방식의 추천시스템이다.

아이템을 검색했을 시에, 하단에 추천 상품으로 (해당 아이템을 본 후에 소비자들이 사는 물건들을 알려준다, 함께 많이 본 아이템을 알려준다.)

미래의 추천 시스템은?

사람에게 있어서 일정한 소비패턴은 존재한다. 하지만 하나의 소비패턴만을 갖고 있지는 않다. 사람에게는 최소 10개 이상의 소비 패턴이 있다. 이를 다 찾아 낸다면 더 정확한 추천이 될 것이다.

추천을 해도 소용이 없는 이유?

사람은 시간에 따라, 기분에 따라, 여러가지 환경적인 요소에 따라서 살 물건도 안 사게 되고, 안살 물건도 사게 된다. 좀 더 세그먼트를 충실히 나눌 필요가 있다고 본다. 그래야 가치가 있다.

물론, 가격적인 측면도 고려해야할 것 같다. 소득 수준, 지불 능력 이러한 거 또한 파악을 해봐야한다.

결론: 제품, 서비스에 대한 도메인 지식과 소비자의 다양한 데이터가 필요하다. 그리고 이를 분석할 기법, 기술의 발달이 요구된다.

사용자 기반 추천시스템보다 아이템 기반 추천시스템을 사용하는 이유?

🡪 비슷한 상품(영화)를 좋아(구매)한다고 해서 사람들의 취향이 비슷하다고 판단하기는 어려운 경우가 많기 때문이다. 매우 유명한 영화는 취향과 관계없이 대부분의 사람들이 관람하는 경우가 많고, 사용자들이 평점을 매긴 상품(영화)의 개수가 많지 않은 경우가 일반적임. 그러므로 이를 기반으로 다른 사람과의 유사도를 비교하기가 어려운 부분이 있다.

머신러닝 분석 흐름

데이터 전처리(인코딩, 피처 스케일링, 정규화) 🡪 학습/테스트 데이터 세트 분리 🡪 교차 검증 🡪 GridSearchCV(교차 검증과 하이퍼 파라미터 튜닝을 한 번에) 🡪 머신러닝 모델 활용해서 예측 🡪 정확도, 오차행렬을 활용한 정밀도와 재현율, F1스코어, ROC곡선과 AUC 평가