Recommender System – 추천 시스템

추천 시스템, 혹은 추천 모델링은 ‘특정 시점’에 ‘특정 고객’이 좋아할 만한 리스트를 찾아주는 것

e-commerce 영역

1. Implicit Score
2. Explicit Score
3. Contents

* 이 중 Contents는 아이템 자체에 대한 정보를 의미

과거 cncjs

뉴스 추천 -> 협업 필터링에 기반한 state of the art에 가까운 알고리즘 사용

State of the art와 비슷한 것 : CF(Collaboration Filtering), DL(Deep Learning)

최신의 알고리즘보다는 훨씬 쉽고 간단하면서도 연산량이 작고, 그럼에도 불구하고 고급 알고리즘들과 비슷한 성능을 낼 수 있는 Association Rule

대부분의 추천 시스템은 대량의 정보를 이용하여 추천하는 B2C비즈니스 모델

추천 시스템에서의 가장 널리 알려진 장애 요소

: Sparsity Problem

* 추천할 아이템과 고객은 계속 늘어남
* 고객이 실제로 소비하게 되는 콘텐츠나 아이템의 비율은 현저하게 감소
* e-commerce 영역 – 인기 상품이나 카테고리는 한정 but, 서비스의 양적 향상은 끝없음
* 소비의 헤게모니 : 오프라인 -> 모바일

영화 추천, 맛집 추천 서비스가 아닌 대부분의 추천 서비스에서 나타나는 문제점

: Information Utilization Problem

* 추천 시스템 구축에 활용하기 위한 데이터, 정보들을 올바르게 활용하기 위한 고민에서 나옴
* E- commerce -> 고객들은 상품을 눌러보고, 다른 상품도 살펴보고, 장바구니에 넣어두고 구매 -> Implicit Score (암묵 점수)
* 로그 데이터 속에 숨어있는 정보를 이용하는 것이 어려움
* 만약 고객의 목록 데이터가 있을 때, 구매가 완료되었다고 상품에 대한 호감을 나타내는 데이터는 아님 -> 환불 또는 교환이 일어났다면 -> Explicit Score ( 명시 점수 : 영화 평점에 대한 rating 같은 점수) 처럼 데이터를 Utilization 하는 과정 필요
* 그렇다고 Explicit Score가 항상 좋은 데이터는 아님
* 대부분 잘 정리된 명시 저수의 경우 Sparsity Problem을 심각하게 겪을 것

과거 추천 시스템에서는 KDD(Knowledge Discover in Database)라는 방법론을 기반으로 알고리즘 개발에 착수

* 데이터셋에서 유의미한 규칙을 찾아내는 것 : Association Rule
* Association Rule은 고객들의 상품 묶음 정보를 규칙으로 표현하는 가장 기본적인 알고리즘, 흔히 장바구니 분석이라고도 불림
* 기초적인 확률론에 기반한 방법으로, 전체 상품중에 고객이 함께 주문한 내역을 살펴본 뒤 상품간의 연관성을 수치화하여 나타내는 알고리즘

추천의 타입은 크게 3가지로 분류

1. 유저의 정보에 기반하여 자동으로 아이템 리스트를 추려주는 Personalized recommender(개인화 추천)
2. rating기반의 인기 상품이나 현재 상품기준(AR, Association Rule)순위 상품을 추천해주는 Non-personalized recommender
3. 이 방법은 주로 Cold Start Problem(개인화 추천 모델링을 위한 유저정보 혹은 아이템 정보가 부족한 상황)이 발생하는 상황이나 개인화추천이 잘 적용되지 않는 추천 영역에 사용된다. 그리고 마지막으로 Attribute-based recommender 방법이 있다. 아이템 자체가 가지고 있는 정보, 즉 Contents 정보를 활용하여 추천하는 방법으로 Cold Start 문제를 해결하는 조금 더 세련된 방법이라고 할 수 있다. 뒤에 설명할 CF(Collaborative Filtering. 이하 CF)와 상호 보완적인 알고리즘인 Content-based approach 라고도 불린다.
4. 위의 세 가지 타입에 매칭되는 대표적인 알고리즘은 Personalized recommender - CF, Non-personalized recommender - AR, Attribute-based recommender - Content based approach 라고 할 수 있다.

통계 기반 모델링에서 사용되는 알고리즘

* 추천 시스템에서 통계 기반 모델링이라는 것은 ‘이상’ 징후를 보이는 아이템을 추출해내는 작업

1. 카이제곱 분포

* χ2 = Σ (관측값 - 기댓값)2 / 기댓값
* 카이제곱 계수를 이용하여, 관측값과 기댓값을 각각 관측 빈도와 기대 빈도로 치환하여 카이제곱 분포에 대한 계수를 계산한 뒤 이를 비정상적인 분포 검정에 활용하는 것

1. 쿨백-라이블러 발산, KLD(Kullback-Leibler divergence)

* 두 확률분포의 차이를 계산
* 어떤 이상적인 분포에 대해, 그 분포를 근사하는 다른 분포를 사용하여 샘플링 한다면 발생할 수 있는 정보 엔트로피 차이 계산
* 상대 엔트로피(relative entropy), 인포메이션 다이버전스(information divergence)라고도 하는데 정보이른의 한 방법
* 잘 일어나지 않는 사건은 자주 일어나는 사건보다 정보량이 높다
* 수식, 알고리즘 간단

CF기반 모델링

1. Memory based 알고리즘

* Neighborhood model의 User based, Item based

1. Model based 알고리즘

* Matrix Factorization(MF), RBM, 베이지안

1. 메모리 기반 알고리즘은 유저와 아이템에 대한 matrix를 만든 뒤, 유저기반 혹은 아이템 기반으로 유사한 객체를 찾은 뒤 빈공간을 추론하는 알고리즘
2. 모델 기반 알고리즘에서 가장 널리 사용되는 MF는 유저나 상품의 정보를 나타내는 벡터를 PCA나 SVD같은 알고리즘으로 분해하거나 축소하는 방법
3. MF의 가장 대표적인 방법은 SVD(Singular Value Decomposition)
4. 특이값 분해라고 하는데, 고유값 분해처럼 행렬을 대각화하여 분해하는 방법 중 하나
5. 고유값 분해와 다른 점은 nXn의 정방행렬이 아니어도 분해가 가능하다는 것
6. 이는 Sparse한 특성을 가지는 추천 시스템에서의 Matrix를 분해하는 것에 안성맞춤
7. SVD를 비롯한 MF에서 목적함수는, Predicted rating을 구하는 Matrix Completion의 경우, 최적의 해(Rating – Predicted Rating의 최소)가 없이 근사값을 추론하는 문제이다. 따라서 Gradient Descent 알고리즘, ALS(Alternating Least Square) 알고리즘 등으로, global minimum에 근접하는 thresh를 선정하여 이를 objective로 삼아 구하는 문제로 볼 수 있다. 일반적으로는 GD가 우수하지만, ALS는 병렬 처리 환경에서 좋은 성능을 보인다고 알려져 있다.
8. 참고 : 학습이 완료된 후 user나 item에 대한 입력값의 행렬 연산 결과를 prediction을 할 때, 예상을 해야하는 결측값(Matrix에서 \*으로 표기된 부분)의 초기값은 binary로 보정하거나 평균 혹은 중앙값으로 보정하기도 한다.
9. 그래서 일반적으로 CF 기반의 추천 시스템을 구축할 때, 가장 많이 사용하는 알고리즘 스택은 SVD + ALS를 기반으로 한 Hybrid 방법이 많다.