4. 하이브리드 시스템

User에게 만족도 높은 추천을 제공하기 위해서는 추천 적용 도메인의 특성에 맞는 알고리즘을 선택하고 User, Item의 다양한 정보를 활용해야 한다.

해당 방식의 출현 배경은 앞선 두 가지 방식(콘텐츠 기반, 협업필터링)의 한계를 보완하기 위해서이다.
- 콘텐츠 기반 접근 방식은 Rating이 없는 Item을 추천할 수 있지만 과도한 특수화 경향이 있다.
- 협업필터링은 높은 seren-dipity를 보여주지만 User가 Rating을 부여하지 않은 Item에 대해서 추천을 하지 않는다.
—> 이로 인해 두 방식의 장점을 극대화하며 단점을 보완하고자 해당 방식이 제안되었다.

Bruke(2002; 2007)에서 제안된 하이브리드 전략은 아래와 같다.
1. 다른 추천 기준을 지닌 여러 알고리즘을 학습한 뒤 추천 점수의 가중 평균합을 구하는 방법
- 여러 알고리즘의 결과를 활용할 수 있지만 각 예측 결과를 정규화 해야 하며 가중치에 대한 정의가 중요하다.
2. 학습된 여러 추천 엔진 중에서 현재 이슈(상황)에 가장 적합한 추천 엔진을 선택하는 방법
- 현재 이슈를 파악하기 위한 추가적인 업무가 필요하다.
3. 각 알고리즘들의 추천 결과를 혼합하여 보여주는 방법
- 추천 결과의 다양성을 높게 보여줄 수 있다.
4. 각 알고리즘에 사용되는 모든 변수를 단일 알고리즘의 변수로 병합하여 이용하는 방법
5. 한 알고리즘이 추천한 Item을 다음 알고리즘의 후보로 이용해 각 단계별로 더 세밀한 추천을 하거나 한 알고리즘의 추천 점수를 다른 알고리즘의 변수로 이용하는 방법
6. 각 알고리즘의 추천 결과를 바탕으로 메타 알고리즘을 학습하는 앙상블 방법

Adomavicius and Tuzhilin(2005)는 아래 네 가지의 하이브리드 전략을 제시하였다.
1. 독립된 추천 결과를 조합(조합 방법에 따라 Bruke의 전략과 유사)
2. 콘텐츠 기반 정보를 협업필터링에 적용하는 것으로 User의 Rating이 아닌 콘텐츠 기반 User 프로파일을 이용하는 방법
3. LSI, PLSI와 같은 알고리즘을 이용하여 협업 필터링의 정보를 콘텐츠 기반 접근 방식에 융합하는 방법
4. 협업 필터링과 콘텐츠 기반 접근방식을 동시에 고려한 단일 모델 구축 방법

McNee et al(2006)은 논문 추천 시스템에서 콘텐츠 기반과 협업필터링에 사용 가능한 알고리즘을 제시하였다.
- 이웃기반 협업필터링을 이용할 경우에는 seren-dipity가 높은 경향이 있으며 나이브 베이지안을 이용할 경우에는 선택한 논문이 많이 참조된 논문을 먼저 추천한다.
- User-Item 행렬을 PLSI(차원축소 기법)로 학습한 후 논문을 추천할 경우 선택한 논문과 분야적으로 매우 유사한 논문들이 우선 추천되었다.
- TF-IDF를 이용하여 콘텐츠 기반 추천을 할 경우 선택한 논문들과 내용이 매우 유사한 논문들이 우선 추천되었다.

Claypool et al(1999)는 적절한 가중치 설정을 위해 초기 가중치를 모두 동일하게 설정한 후 추천 결과에 대한 User 피드백 정보를 활용하여 설명력이 가장 좋은 가중치를 실시간으로 조절하는 방법을 제안하였다.

Fan et al.(2014)는 각 추천 시스템 방식을 가중치를 달리하여 선형 결합할 경우 추천 만족도가 개선되는 것을 확인했다.

Kim et al.(2002)는 두 가지 방식을 동시에 사용하는 것이 아니라 구매이력이 존재하는 User에 대해서는 협력필터링을 통해 추천하고 그렇지 않은 User에 대해서는 사용자 프로필을 사용하여 유사도를 측정한다.

Melvile et al.(2002)는 Item을 선택한 User 정보를 통해 Item간의 유사도를 계산하였다. 이를 통해 User-Item 행렬에 각 User가 선택한 Item과 유사한 다른 Item을 가상으로 선택한 것으로 여기는 유사 User 벡터를 정의했다. 밀도가 높아진 User-Item 행렬을 이용하여 이웃기반 협업 필터링을 수행하였다.

Goldberg et al.(2001)은 Sparsity를 해결하기 위해 PCA를 수행하고 클러스터링을 수행하여 유사한 이웃을 찾았다.

Jeh and Widom(2002)는 SimRank를 고안했는데 User와 Item의 이분 그래프를 구축하여 유사도를 정의한다. 두 User의 유사도는 각 User가 선택한 Item간의 평균 유사도로 정의되고 Item간의 유사도는 User간의 평균 유사도로 회귀적 정의된다. 따라서 SimRank를 이용할 경우 동일한 Item을 선택하지 않아도 유사한 Item을 선택할 경우 두 User는 높은 유사도를 지니게 된다.

Sawant(2013)는 이분 그래프에 클러스터링을 수행한 뒤 random walks model을 적용하여 개인 User 혹은 클러스터 간의 영향도를 계산하여 협업필터링을 수행함으로써 데이터 희소성 문제 해결과 User간의 추천 영향력을 계산했다.

Vozalis and Margaritis(2004)는 이웃기반 협업 필터링으로 유사 이웃 후보군을 선택한 뒤 인구통계학 정보를 이용하여 추천 대상과 유사한 User들의 정보만을 이용하는 연쇄방법을 제안하였다.

Chow et al.(2014)는 이분 그래프에 사용자의 선호 정보를 결합하여 Personalized PageRank를 적용한 개인화 추천 알고리즘을 고안했다.

Basilico and Hofmann(2004)는 User - Item의 특성 변수들을 종합하여 User - Item의 유사도 커널을 정의하는 프레임을 제안했다.

Ganu et al.(2009)은 Item에 작성된 리뷰로부터 추출된 토픽과 의미 벡터를 이용하여 이웃기반 협업필터리을 보강했다.

McAuley and Leskovec(2013)는 리뷰로부터 토픽을 추출할 경우 LDA와 같은 기법은 Rating을 이용하지 않는 점을 보강하기 위해 리뷰와 Rating을 모두 이용해 토픽을 추출하는 Hidden Factors as Topics(HFT)를 제안하였다.

Ling et al.(2014)는 User들이 선택하는 Item에는 잠재 요인이 있다고 가정하고 해당 요인을 리뷰로부터 추출하였다.

5. 연관성분석

- 연관성 분석 = 장바구니 분석
- 핵심은 조건부 확률로써 “사건 A가 발생했을 때 사건 B가 발생하는 것”을 의미한다.
- 즉, User가 A를 구매했을 때 B에 대해서 만족한다와 같은 의미
- A의 정보로써 인구통계학적 정보를 사용할 수 있고 B는 추천 대상 User에게 추천할 항목으로 정의할 수 있다.

*** 판단 기준 ***
- 지지도 : 전체 발생 빈도수
- 신뢰도 : 조건부 확률을 통해 항목간 관련 정도

- 동시 출현 빈도수를 기반으로 추천 항목의 우선순위가 정해지므로 데이터 희소성과 Grey Sheep 문제를 해결할 수 있다.
- 하지만 Item과 User의 수가 늘어날 경우 데이터 희소성 문제가 발생해 연산이 복잡해지고 계산량이 많아진다.

- Jin et al.(2010)는 RFM 기법을 사용해 고객을 segmentation 한 후 연관성 분석을 수행하여 교차판매 전략을 수립하였다.