

추천시스템 논문 리뷰

제목 : 추천 시스템 기법 연구동향 분석(Review and Analysis of Recommender Systems)

1. 콘텐츠 기반 접근 방식

- 콘텐츠 기반 접근 방식은 정보검색기술에 바탕을 두고 있다.
- Item의 Contents를 직접 분석해서 Item과 Item, Item과 User 선호도간 유사성을 분석하여 추천하는 방식이다.
- User가 입력한 프로파일 정보나 과거 구매 내역을 활용해 Item의 선호도를 파악한 후 Item Category와 User가 선호하는 Item 간의 Similarity를 측정한다. 이후 Similarity가 가장 높은 Item을 추천하게 된다.

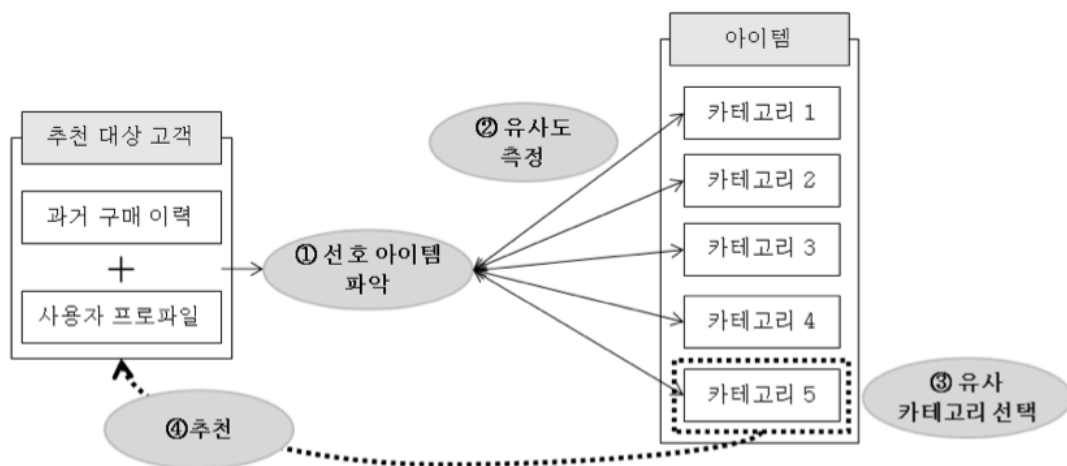


Figure 2. The algorithm of content-based approach

*** 콘텐츠 기반 접근 방식의 특징은 크게 다음과 같다.

- (1) 독립적 정보 활용 : User의 구매 이력이나 프로파일 정보만 존재하면 된다. 추천 대상 User의 독립적인 정보만이 필요하므로 정보가 부족해도 유용하게 쓰일 수 있다. 때문에 대부분의 Item에 대해 평가점수가 존재하거나 User의 과거 구매 이력이 충분해야 성능을 보장할 수 있다.
- (2) 새로운 Item : 일반적으로 새로운 Item이 추가되는 경우 추천리스트에 포함되지 않을 수 있다. 이를 First rater문제라고 하는데 해당 방식은 Item간의 속성을 사용하기 때문에 First rater 문제에서 자유롭다.
- (3) 과도한 특수화 : User의 과거 구매 이력을 사용하기 때문에 이전에 구매한 Item과 유사한 상품들만 추천을 하게 되어 추천 상품의 다양성을 보장할 수 없다.
- (4) 아이템 속성 분석 : 구조적 및 비구조적 데이터를 활용하여 Item이 가지는 콘텐츠의 유사도를 구하는데 이 때 코사인 유사도를 이용한다.

- (5) 사용자 선호도 학습 : User의 선호도를 파악하기 위해 사용되는 첫번째 방법은 User가 직접 선호도를 입력하는 것이다. 체크상자 혹은 텍스트 형식으로 기재를 할 수 있도록 하여 프로파일 정보를 구성한다. 두 번째 방법은 과거 이력을 활용하는 것이다.

2. 협력 필터링

- 협력 필터링은 현재(2015)까지 알려진 가장 우수한 성능의 기법이다.
- 특정 Item에 대해 선호도가 유사한 User들은 다른 Item에 대해서도 비슷한 선호도를 보일 것이다.
- User가 Item에 대해 평가한 정보를 사용해 선호도를 예측한다. -> 즉, 추천 대상인 User와 취향이 비슷한 User를 선정하고 그들이 선호하는 Item을 추천하기 때문에 Item의 다양성이 보장된다.
- 협력 필터링은 크게 기억기반 협력 필터링과 모델 기반 협력 필터링으로 나뉜다.
- (1) 기억 기반 협력 필터링 : User간 Similarity를 계산한 후 Similarity가 높은 User가 구매한 Item을 추천해주는 방식
- (2) 모델 기반 협력 필터링 : 기억 기반 협력 필터링 + 군집화, 분류, 예측 단계에서 기계학습 및 데이터 마이닝 적용

*** 협력 필터링의 특징은 다음과 같다.

- (1) 데이터 희소성(Sparsity) : 협력 필터링에서는 선호도 예측을 위해 아래와 같이 cross-tab 형식으로 행렬을 만든다.

Table 3. Example of ‘user-item’ matrix

	아이템 1	아이템 2	아이템 3	아이템 4
사용자 a	4	?	5	
사용자 b	4	2	1	
사용자 c	3		2	
사용자 d	4	4		
사용자 e				

이 때, 아직 평가 받지 못한(혹은 새로운) Item은 예측이 거의 불가능한데 이를 Cold Start 문제라고 한다. 또한 앞서 콘텐츠 기반 필터링에서 First rater 문제에 직면할 수 있다.

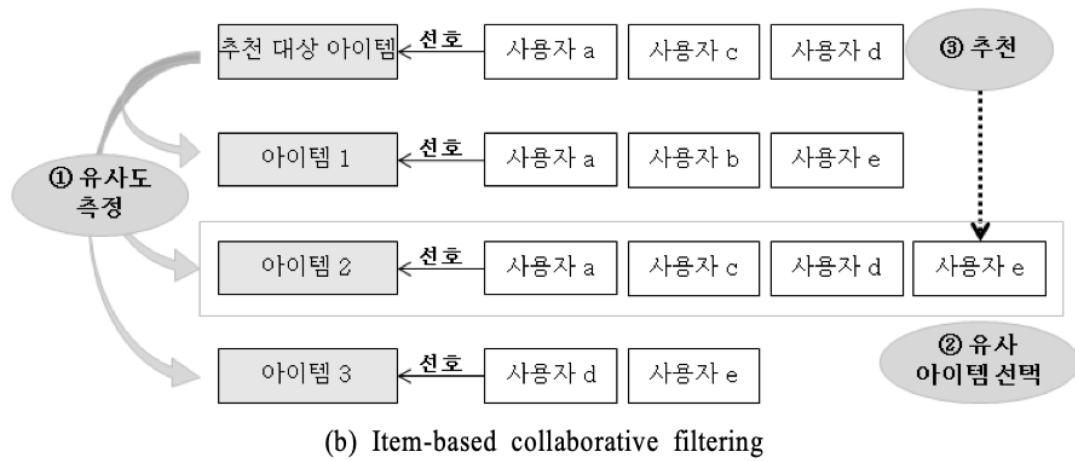
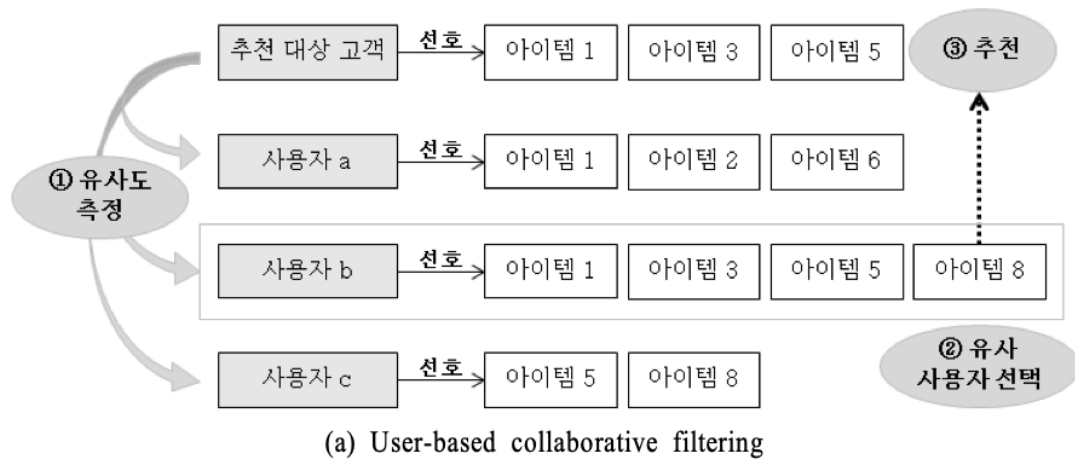
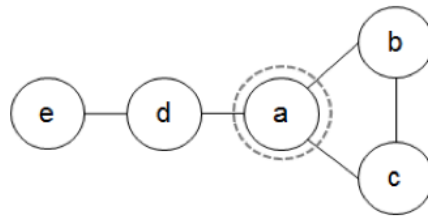


Figure 3. The algorithm of collaborative filtering

이를 해결하기 위해 협력필터링 기법 + 사회연결망 기법이 시도되었다. 기존 User들 간의 네트워크를 생성한 후 중심성이 높은 User를 이웃으로 정하고 그들이 선택한 Item을 추천하는 방식이다. 중심성이 높은 고객이 선택한 Item을 추천하면 User들간의 비슷한 선호도를 반영할 수 있어 대중성과 신뢰성을 확보할 수 있다는 가정이 깔려 있다.

	사용자 a	사용자 b	사용자 c	사용자 d	사용자 e
사용자 a	1	0.7	0.8	0.8	0
사용자 b	0.7	1	0.6	0	0
사용자 c	0.4	0.6	1	0	0
사용자 d	0.8	0	0	1	0.9
사용자 e	0	0	0	0.9	1

(a) Similarity matrix



(b) User network

Figure 4. Example of recommendation for new customers using collaborative filtering and social network analysis

위 그림의 경우는 User a가 중심성이 높기 때문에 해당 User가 구매한 Item을 신규 User에게 추천한다.

Kim et al(2003)에서는 신규 User의 과거 구매이력이 없고 프로필 정보가 있을 경우 이를 기반으로 User간의 유사도로 선호도가 유사한 User들을 탐색하고 이들의 선호도를 통해 신규 User의 선호도를 예측했다. 해당 방식은 과거 구매 이력이 없는 User에 대해서도 개인화된 추천목록을 제공할 수 있다는 장점이 있지만 프로필 정보가 필수적이다.

Ko et al(2014)는 User를 “주요 사용자”와 “일반 사용자”로 분류한 방법을 제안했는데 주요 사용자는 많은 Item을 평가했거나 구매한 User로 User - Item Cross Table에 결측치가 거의 없고 데이터의 Sparsity를 해결할 수 있다. 추천 과정은 아래와 같다.

- 1. “주요 사용자”의 데이터로부터 선호하는 Item간 유사성을 계산하여 Item을 군집화 한다.
- 2. 1과 같은 방식으로 “일반 사용자”를 군집화 한다.
- 3. “주요 사용자”와 “일반 사용자”의 유사한 군집끼리 매칭킨다.
- 4. “일반 사용자”군집에만 없는 Item을 추천한다.

추가적으로 데이터의 Sparsity를 근본적으로 해결하기 위해 차원축소 방식이 연구되었는데 SVD(특이값 분해)가 대표적이다.

- (2) 확장성(Scalability) : 최근 Item의 종류가 다양해지고 User가 늘어남에 따라 데이터가 커지고 있으며 추천 알고리즘 구현 시 계산량이 급증하여 많은 시간과 비용이 소모된다. Park et al(2014)은 유사도를 기준으로 User혹은 Item을 N만큼 추출하여 선호도를 예측한다. 사전에 필터링을 한 후에 선호도를 예측하기 때문에 계산 속도와 양 측면에서 기존의 협력 필터링 기법 보다 뛰어나다.

- (3) Grey sheep : 일관성 없는 의견(=선호도)을 가지는 User들의 데이터는 전체 선호 경향과 패턴을 파악하는데 방해가 되며 예측 성능을 저하시킨다. Claypool et al(1999)는 콘텐츠 기반 방식과 협력 필터링 방식을 모두 사용하여 가중 평균 값을 예측값으로 사용하는 방법을 제안했다. 두 방식의 각각 예측값의 차이가 큰 User의 경우 가중치를 낮게 주어 예측 오차를 줄이는 방식이다.

-(4) Shilling Attack : 판매자 입장에서 본인의 Item이 많이 팔리고 경쟁사의 Item이 적게 팔리게 하기 위해 Rating을 임의로 평가할 수 있다. 이처럼 악의적으로 Rating을 하여 추천 시스템이 정상적으로 실현되지 못하게 하는 것을 Shilling Attack이라고 한다.