개별 평가 편향 요소 경감을 통한 협업 필터링의 개선

박경화[1], 김병희[2], 박태서[3], 장병탁[1,2,3]
[1]서울대학교 뇌과학 협동과정
[2]서울대학교 컴퓨터공학부
[3]서울대학교 인지과학 협동과정
{kwpark, bhkim, tspark, btzhang}@bi.snu.ac.kr

Enhancing Collaborative Filtering by Reducing Rating Biases

Kyung-Wha Park[1], Byoung-Hee Kim[2], Tae-Suh Park[3], Byoung-Tak Zhang[1,2,3]
[1]Brain Science Program, Seoul National University
[2]Department of Computer Science and Engineering, Seoul National University
[3]Cognitive Science Program, Seoul National University

요 약

대표적인 추천 기법으로 널리 사용되고 있는 협업 필터링의 적용에 있어 사용자 평점 데이터의 무결성은 필수적인 전제이다. 사용자의 평점 부여 행동에 내재된 편향 요소로 인해 발생하는 몇 가지 특징적인 평점 패턴을 제시하며, 이를 기반으로 편향 요소를 경감시키는 기법을 제안한다. 대표적인 영화 추천 벤치마크 데이터인 MovieLens 데이터를 이용하여 제안한 기법의 적용 전·후 다양한 협업 필터링 기법의 성능을 분석한 결과, 평점 예측 오류가 감소하는 경향을 확인하였다.

1. 서 론

추천 시스템은 정보 홍수 시대에 음악, 책, 영화 등의 서비스를 위한 필수 도구가 되고 있다. 추천 시스템의 자동화 및 성능 향상을 위한 연구가 데이터 마이닝, 마케팅, 인지 과학, 인공지능 등의 분야에서 매우 활발하게 진행되고 있다.

추천 시스템은 사용자 취향에 맞는 개인화된 추천을 제공하기 위해 상품 혹은 아이템에 대해 가지는 사용자 관심(user interest)의 패턴을 분석한다[1]. 사용자 관심을 분석하기 위해 별 5개 평점과 같이 주로 외적으로 드러나는 평점을 사용한다. 이 평점들을 모은 데이터베이스를 통해 기계학습과 통계적 패턴 마이닝을 수행하여 추천 시스템을 설계하게 된다.

평점 데이터는 사용자 편향, 아이템 편향, 그리고 그 두 개의 관계에 의해 발생한 내재적인 편향의 세 가지 편향 요소를 포함하고 있다. 예를 들어, 비판적인 사용자는 보통의 사용자에 비해 낮은 점수를 자주 주고, 평판이 좋은 아이템은 다른 비슷한 아이템에 비해 보다 높은 평점을 받기 쉽다[2]. Koenigstein에 따르면, 비록 추천 시스템이 좋은 설명력을 가진 개인화된 성능을 가졌더라도, 대부분의 시스템은 위의 세 가지 편향을 간과하고 있다[3]. 진정한 개인화와 빠른 개별 문맥 적응을 위해, 이러한 편향성에 대한 체계적인 분석이 필요하다.

평점에서의 편향에 대해 추천 시스템 분야에서 다양한 연구가 진행되었다. Pennock 등은 협업 필터링에 아이템에 대한 내적 선호도를 표현하고 점수를 매긴 모든 아이템의 "진짜" 평점에 대한 확률 기반 모델을 제안했다[4]. Pennock은 모든 편향 요소를 가우시안 노이즈로 설명했다. Shan과 Banerjee는 사용자 편향과 아이템 편향을 residual approach로 모델링하고, 모델에 따라 편향 요소를 제거한 후 은닉 요소 기반 협업 필터링 기법을 적용할 것을 제안했다[2]. 이 모델도 마찬가지로 사용자와 아이템 각각의 편향을 가우시안 노이즈로 모델링했다.

Koenigstein은 기존의 Shan과 Banerjee의 제안이 개인화된 추천 아이템 순위에 결과를 주지 않는다고 주장했다. 그러면서 개인 특성에 관련이 있는 편향과 관련 없는 편향을 구분했고, 아이템 분류, 사용자의 연속적 세션 안에서의 평점, 그리고 아이템의 temporal dynamics와 같이 개인화 목적에 관련 없는 요소를 제거하는 모델을 제안했다. 이 연구에서 평점의 실제 특성이 간격 척도 혹은 서열 척도일 지라도, 연구자들은 평점이 비율 척도라고 가정했다[5]. 평점을 추천 시스템의 통계적 도구로서 올바르게 적용하려면, 간격 척도 혹은 서열 척도로 리스케일링이 필요하다.

본 논문에서는 관측된 평점은 사용자의 내적 선호도가 평가시 외적으로 드러나는 유형에 따른 왜곡·편향된 결과로 설명하고자 한다. 편향에 대한 모델로서 기존의 정규분포 기반이 아닌 액티브 사용자의 평점 패턴을 관찰하여 발견한 여섯 개의 패턴을 RBIAS라는 이름으로 제안한다. 사용자 편향을 이들 여섯 개의 유형을 기반으로 하여 모델링 하고, 사용자 편향을 줄이는 것으로 사용자 선호도를 정확히 찾아낼 수 있게 되므로, 영화 추천 데이터 셋에 적용시켜 추천 성능이 향상되는 것을 보인다.

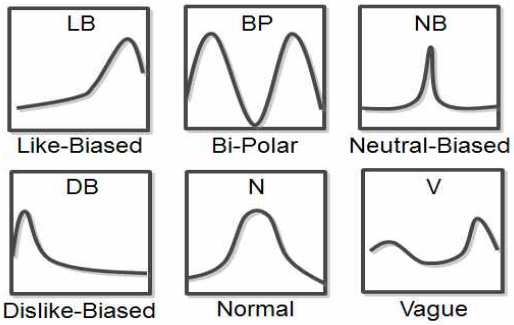

그림 1. RBIAS: 6가지 평가 유형 패턴의 개념도 (X축은 ordinal 혹은 interval scale에서의 평점, Y축은 평점의 빈도를 나타냄)

2. 편향 요소 경감 기법

2.1 RBIAS : 평가 유형 모델링

외적으로 드러나는 평점은 추천 시스템에 가입한 액티브 사용자들에 의해 매겨지고 대부분은 평점을 주고 싶은 영화를 골라서 평점을 주게 된다. 사용자가 평점을 주고 싶은 영화를 골라서 점수를 주는 상황 설정에서 다양한 종류의 편향에 노출된다. 사용자 인터페이스, 사용자의 감정 상태, 또는 아이템의 속성에 영향을 받는다. 그러나 이러한 편향들은 숨겨져 있고, 편향을 특정 짓기 위해 관측된 패턴을 먼저 분석할 필요가 있다. 관측 결과 특징적인 패턴을 6가지 유형으로 구분할 수 있었으며 (그림 1), 각 유형은 다음과 같다: 1) 높은 점수를 주로 주는 like-biased (LB), 2) 낮은 점수를 주로 주는 dislike-biased (DB), 3) 중간 점수가 다수를 차지하는 neutral-biased (NB), 4) 중간 점수보다 양 극단의 점수를 주는 bipolar (BP), 5) 정규 분포를 따르는 normal (N), 그리고 6) 패턴이 특정되지 않는 vague (V) 유형.

6가지 유형을 결정짓는 규칙은 표 1과 같다. 한 사용자가 매긴 평점들을 모아서 먼저 정규성을 확인한다(본 논문에서는 Liliefors test를 적용). 여기서 정규성 귀무가설을 기각하지 못하면 이 사용자는 normal 유형으로 결정한다. 통과하지 못한다면 표 1에 적힌 결정 규칙에 따라 순서대로 검증을 하고 여기서도 통과하지 못한다면 vague 유형으로 결정하게 된다.

2.2 리스케일링 기법을 통한 편향 경감

기존의 추천 시스템에서는 보통 사용자의 평가 유형을 구분하지 않고 일괄적으로 4점 이상을 매긴 영화만 선호하는 것으로 단순히 가정했었다. 그러나 사용자 간의 평가 유형이 차이가 난다는 분석 결과에 따르면 4점 이상으로 선호를 일괄적으로 판단하는 것은 적절하지 못하다. 리스케일링을 통해 3점을 중심으로 편향 요소를 교정함으로서 선호를 특정 점수 기준으로 판별하는 것이 적절하다는 근거가 생긴다.

표 1. RBIAS 결정 규칙 (* 표시는 'don't care'를 의미)

	LB	DB	BP	NB
최빈값	>= 3.5	<= 2.5	*	>= 2.5 & <= 3.5
평균	>= 3.5	<= 2.5	>= 2.5 & <= 3.5	*
표준편차	*	*	>= 1.43	< 1.3
왜도	> 0	< 0	*	*

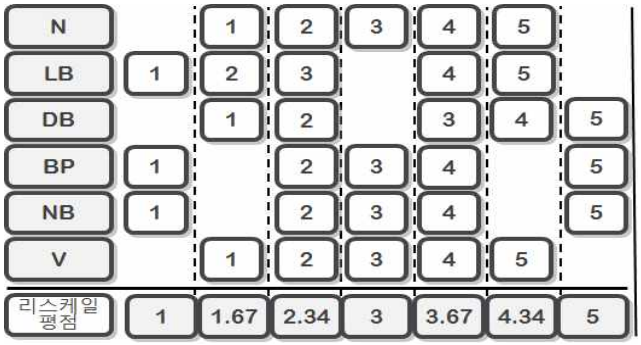

그림 2. 5점 척도에서 평가 유형에 따른 편향을 줄이는 리스케일링 모델

리스케일링 기법은 사용자가 투고한 평점으로 실제 선호도를 나타내는 점수로 변환을 시도하는 것으로, 평점에 숨겨진 편향 요소의 경감을 시도했다. 예를 들어 like-biased 유형의 5점은 dislike-biased 유형의 5점을 비교했을 때 dislike-biased 유형의 5점이 like-biased 유형보다 더 비중이 크다는 것을 나타내고자 했다. Dislike-biased 유형은 낮은 점수 쪽으로 편향되어 있기에 이것을 교정하기 위해 투고한 점수에 규칙에 맞는 점수를 더하는 방식으로 변환을 시행했다. 이를 통해 명확한 간격 척도로 변환하는 모델을 그림 2에 나타내고 있다. 예를 들어, 본래 LB의 2점은 리스케일링을 통해 1.67점으로 감소시켜 변환된다.

3. 영화 추천에서의 협업 필터링 개선

3.1 데이터셋

영화 추천에 널리 쓰이는 5점 척도로 매겨진 MovieLens-100K 데이터 셋(이하 ML-100K)을 사용했다. ML-100K 셋은 MovieLens 웹사이트에서 1997년 9월부터 1998년 4월까지 수집된 데이터이며, 943명의 사용자가 1,682편의 영화에 대해 부여한 10만 건의 평점 정보를 포함하고 있다. 해당 웹사이트에 사용자들이 로그인하여 5점 (꼭 봐야 하는 영화)에서 1점 (최악인 영화)까지 평점을 줄 영화를 선택한다.

3.2 분석 결과

ML-100K 의 사용자들은 절대 다수가 LB 유형 (53%), 그 다음이 NB 유형 (22%), V 유형 (18%), N 유형 (3.5%), BP 유형 (2%), 제일 수가 적은 DB 유형 (1.5%)으로 구분되었다.

RBIAS를 ML-100K에 적용한 결과는 그림 3과 같다. 모델이 구분한 평가 유형 별로 투고한 평점의 분포를 나타낸다. N은 정규 분포를 따르는 모습을 보이고 있으며, LB는 4점과 5점에 투고의 50% 이상이 몰려있고 DB는 1점과 2점에 50% 이상이 몰려있다. 이로서 유저의 평점 패턴이 존재하며, RBIAS 모델이 잘 대응하고 있다는 것을 보여준다.

ML-100K 데이터 셋을 리스케일링을 통해 변환한 데이

표 2. MovieLens 100K 데이터 셋을 리스케일링하기 전(R)과 후(R')의 주요 추천 기법별 성능 비교

추천 시스템	User-Based CF		Item-Based CF		NMF		BPMF	
	R	R'	R	R'	R	R'	R	R'
Precision	0.77	0.77	0.78	0.78	0.74	0.74	0.75	0.75
Recall	0.97	0.95	0.96	0.95	0.95	0.93	0.97	0.95
F1 measure	0.86	0.85	0.86	0.85	0.83	0.83	0.85	0.84
MAE	0.67	0.58	0.67	0.59	0.74	0.64	0.71	0.61
RMSE	0.87	0.77	0.86	0.78	0.96	0.85	0.90	0.79
HLU	0.78	0.76	0.79	0.77	0.74	0.73	0.76	0.74
NDCG	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94

터를 가지고 여러 협업 필터링 기법의 성능이 어떻게 변하는지 분석했다(표 2). 비교를 위해 두 가지 추천 모드를 검증했다. 첫째는 평점이 3을 넘으면 선호, 넘지 못하면 비선호로 이진화한 선호도를 예측하는 것이다. 여기서 일반적인 지표인 precision, recall, F1 measure를 비교했다. 둘째는 회귀 분석 기반으로 평점을 예측하는 방식이다. 여기서는 MAE (mean absolute error)와 RMSE (root mean squared error)를 비교했다. 추가적인 분석을 위해 랭크 기반 평가 지표인 HLU (half-life utility)와 NDCG (normalized discounted cumulative gain)를 추가했다. 협업 필터링 기법 적용 및 평가는 PREA 추천 툴킷[6]을 이용해 진행했다.

실험 결과, 리스케일링을 한 경우 리스케일링을 하지 않았을 때에 비해 Precision, F1 measure, HLU 는 유의미한 차이가 나지 않지만 NDCG, RMSE, MAE 지표를 기준으로 볼 때 통계적으로 유의미한($p<0.05$) 성능 향상을 확인했다. MAE와 RMSE가 낮아진 것은 서로 다른 사용자의 평가 유형이 하나의 공통된 척도로 변환됐다는 것을 의미한다. 5점 척도보다 조밀한 7점 척도로 바뀐 것도 오차를 낮추는 요인이 된다. NDCG가 향상된 점은 선호도를 파악하는 성능이 향상되어 랭크를 보다 정확히 파악할 수 있다는 것을 의미한다.

4. 결론 및 향후 연구

본 논문에서는 기존 추천 시스템에서 사용의 선호를 일괄적으로 분석하는 데서 더 나아가, 개별 선호도를 판별하는데 근거를 마련하기 위해 사용자의 평가 유형을 관찰하여 특징적인 normal, like-biased, dislike-biased, bipolar, neutral-biased, vague의 6가지 유형으로 분류한 RBIAS 모델을 제안했고, RBIAS 기반으로 투고한 평점을 리스케일링 기법을 통해 내재된 편향을 교정하여 추천 시스템의 성능이 향상 된다는 것을 보여줬다.

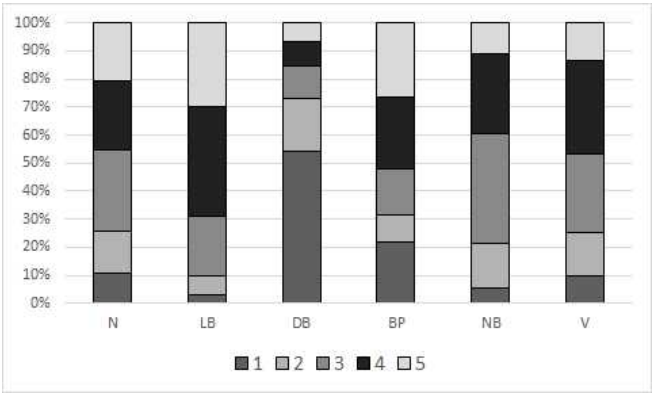

그림 3. MovieLens 100K에서 관찰된 평가 유형 별 평점의 분포

본 논문에서는 MovieLens-100K 데이터 셋을 주로 사용했지만 이 외에 공개된 다른 데이터 셋에서도 실험을 진행 중에 있다. 또한 사용한 데이터 셋은 5점 척도인데 이 외에 7점 척도 혹은 10점 척도에도 실험을 계획 중이며, 영화 외에도 책이나 음악 등 다른 아이템 분야에도 적용하는 실험을 준비 중이다. 또한 RBIAS 모델의 응용으로서 현재는 사용자에 초점을 맞췄지만, 아이템에 초점을 맞춘 실험을 진행 중에 있다.

감사의 글

이 논문은 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구이며(NRF-2010-0017734-Videome), 정부(산업통상자원부)의 재원으로 한국산업기술평가관리원 지원(KEIT-10035348-mLife, KEIT-10044009)을 일부 받았음.

참고문헌

[1] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, **42**(8):30-37, 2009.

[2] H. Shan and A. Banerjee, "Generalized Probabilistic Matrix Factorizations for Collaborative Filtering," in *Proceedings of the IEEE International Conference on Data Mining*, pp. 1025-1030, 2010.

[3] N. Koenigstein, G. Dror, and Y. Koren, "Yahoo! Music Recommendations: Modeling Music Ratings with Temporal Dynamics and Item Taxonomy," in *Proceedings of the Fifth ACM Conference on Recommender Systems*, pp. 165-172, 2011.

[4] D. M. Pennock, E. Horvitz, S. Lawrence, and C. L. Giles, "Collaborative Filtering by Personality Diagnosis : A Hybrid Memory- and Model-Based Approach," in *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, pp. 473-480, 2000.

[5] S. S. Stevens, "On the Theory of Scales of Measurement," *Science*, **103**(2684):677-680, 1946.

[6] J. Lee, M. Sun, and G. Lebanon, "PREA: Personalized Recommendation Algorithms Toolkit," *Journal of Machine Learning Research*, **13**(1):2699-2703, 2012.