협력적 필터링에서 MAE 변화에 관한 연구

이희춘*, 이석준**, 김선옥***

*상지대학교 컴퓨터데이터정보학과, **상지대학교 경영정보학과, ***한라대학교 정보통신공학부

A Study on Changing the MAE in Collaborative Filtering

Lee Hee-choon, Lee Seok-Jun, Kim Sun-Ok

Sang-ji University, Sang-ji University, Halla University

E-mail : choolee@sangji.ac.kr, crco909@yahoo.co.kr, sokim@halla.ac.kr

요 약

협력적 필터링을 이용한 추천시스템은 인터넷 기반 전자상거래에서 좋은 추천 도구로 사용되고 있다. 협력적 필터링 방식은 고객의 선호도를 조사하여 이를 바탕으로 이웃 고객을 선정하고 이들에 대한 선호도를 수집하여 고객이 좋아할 만한 상품을 추천하는 기법이다. 이웃 고객에 대한 정보를 이용하여 추천에 사용하므로 이웃고객이 적은 경우 추천시스템의 예측에 어려움이 생긴다. 본 논문은 추천시스템의 예측 정확도를 높이기 위한 방법으로 희소성이 있는 상품을 우선 선정하고 그들 상품에 대한 선호도를 조사하였다. 그리고 이들에 대한 선호를 나타낸 고객들을 선별하여 추천시스템의 예측 정확도를 향상시키는 방법을 제안한다.

1. 서론

인터넷이 발달됨에 따라 인터넷을 이용한 전자상거래가 활발하게 전개되고 있다. 이에 따라 정보의 양도 많아지고 있으며, 지속적으로 늘어나는 정보의 홍수 속에서 고객은 원하는 서비스와 상품에 대한 정보를 얻기 위한 많은 시간과 노력이 필요하게 되었다. 추천시스템은 고객의 특성에 따라 적절한 상품을 추천함으로써 많은 정보로 인해 결정이 어려운 문제를 해결하기 위한 방안을 제시하였다.

추천시스템은 연관규칙, 특정치 분해 기법 등 여러 가지 방법으로 구현될 수 있는데, 추천시스템의 주요 기법중의 하나는 다양한 정보 속에서 고객에게 적절한 정보를 찾아내는 여과기법이다. 여과기법은 크게 내용기반 필터링과 협력적 필터링으로 나눌 수 있다. 내용기반 필터링은 추천대상 고객이 선호했던 과거의 정보를 이용하여 상품을 추천한다. 협력적 필터링은 고객에 대한 기본 정보와 고객과 유사정도가 있는 이웃고객의 상품 선호도를 이용하여 선호 상품을 추천한다.

협력적 필터링을 이용한 추천시스템은 가장 성공적인 기법으로 인터넷에서 상용화되는 시스템이다. 하지만 다른 고객의 선호도에 대한 정보를 이용하므로 어느 정도의 평가치가 있어야 한다. 적은 평가치를 사용하여 협력적 필터링을 이용한 추천은 추천시스템의 신뢰에 심각한 문제를 일으킬 수 있다. 이것을 추천시스템의 희소성 문제라 하며 이를 해결하기위한 지속적인 연구가 진행되고 있다.

Soboroff[6]는 행렬을 이용한 SVD (Singular Value Decomposition)를 계산하여 희소성의 문제에 접근하여 추천시스템의 계산속도 향상에 기여하였으나 결과적으로 정확도는 크게 나아지지 않았다. Kim[7]은 희소성의 수에 따라 집단을 분리

하여 희소성이 MAE에 미치는 변화를 분석하였고, 분류된 집단에 따라 MAE의 유의적인 차이가 있음을 밝혔다. Pazzani[3]는 데이터를 희소성에 따라 우선 선별하고 선별된 데이터를 속성별로 추출하여 추천시스템의 선호도 예측을 향상시키는 연구를 하였다. 또한 Kim[4]은 희소성이 높은 데이터를 희소하지 않는 상태로 변형하는 데이터 변형기법을 제안하였다. 이 논문에서 사용한 데이터 변형기법은 아이템의 추가 속성 정보에 대한 확률분포를 이용하여 희소성의 데이터를 변경하고, 변경된 선호도 데이터를 협력적 필터링을 이용하여 추천의 성능을 향상 시키는 것이다. 여기서는 다양한 형태의 선호도 평가 값에 대한 데이터들의 특성을 무시하고 확률분포만을 사용하였으므로 각 데이터들에 대한 정보가 정확하게 반영되지 않았다. Melville[5]는 희소성이 있는 사용자의 평가 행렬을 내용기반 필터링을 통해 사용자 평가 행렬을 생성하고, 이를 기반으로 협력적 필터링을 이용하여 추천에 사용하였다. 이 연구에서는 희소성의 문제는 조금 완화되었지만 추천의 정확도는 크게 향상되지 못하였다.

본 논문에서는 희소성의 문제를 해결하기 위해 데이터를 희소성에 따라 우선 선별하고 선별된 상품을 선호한 고객을 조사하여 추천시스템의 예측 성능을 개선하기 위한 방법을 제시하였다.

2. 협력적 필터링 기법

추천시스템에서 협력적 필터링은 널리 사용되는 기법으로 미네소타대학의 GroupLens에서 뉴스 기사의 선정에 이용되었으며, MovieLens에서는 영화 추천을 위한 도구로, Ringo에서는 음악 추천으로 사용되는 등 여러 영역에 적용되었다.

협력적 필터링에서 가장 일반화된 알고리즘은 이웃기반 협력적 필터링이다. 이웃기반 기법은 추천대상고객의 선호도 평가와 더불어 다른 고객의 선호도 평가를 사용하여 상품에 대한 선호도를 예측하는 알고리즘이다. 추천을 위한 선호도 예측을 하기 위하여 우선 추천 대상 고객의 이웃을 선정하여야 한다. 많은 고객 중에서 추천대상 고객의 이웃선정에 대한 다양한 방법이 연구되고 있고, 그중 본 논문에서 사용하는 방법은 추천 대상 고객의 상품에 선호도를 평가한 고객만을 선택하여 이웃으로 선정한다. 선정된 이웃 고객과 추천대상 고객의 유사정도를 알기위해 상관계수를 사용한다. 다음 식은 두 고객과의 유사정도를 나타내는 상관계수 중에서 피어슨 상관계수에 대한 정의이다[8].

$$r_{uj} = \frac{\sum_{i=1}^{m}(R_{u,i} - \overline{R_u})(R_{j,i} - \overline{R_j})}{\sqrt{\sum_{i=1}^{m}(R_{u,i} - \overline{R_u})^2 \cdot \sum_{i=1}^{m}(R_{j,i} - \overline{R_j})^2}} \quad (1)$$

여기에서, r_{uj}는 추천대상 고객 u와 이웃고객 j와의 유사정도를 나타내는 가중치이며, $R_{u,i}$는 추천대상 고객 u가 평가한 상품i 에 대한 선호도 평가치이고, $R_{j,i}$는 이웃고객 j가 평가한 상품i 에 대한 선호도 평가이다. $\overline{R_u}$는 추천 대상 고객 u가 평가한 모든 상품들에 대한 평균이고, $\overline{R_j}$는 추천대상 고객의 이웃인 j고객의 상품 선호도평가에 대한 평균값이다. 유사도 가중치를 계산하기 위해 사용되는 평가치는 추천 대상 고객 u와 이웃고객 j가 공통으로 평가한 상품의 평가치만 사용한다.

다음으로 추천 대상 고객에게 추천 상품에 대한 선호도 예측 값을 선정해야 한다. 상품에 대한 선호도 예측은 추천 대상 고객의 평균과 추천대상 고객의 이웃들이 평가한 평가 값 그리고 이들 이웃고객의 평균값을 사용하며, 식(1)에서 소개된 이웃의 유사도 가중치를 알아야한다. 다음 식은 협력적 필터링을 사용하여 선호도 예측을 계산하기 위한 알고리즘이다[9].

$$\widehat{U_x} = \overline{U} + \frac{\sum_{J \in Raters}(J_x - \overline{J})r_{uj}}{\sum_{J \in Raters}|r_{uj}|},\ \overline{J} = \frac{\sum_{i=1}^{n} J_i}{n},\ i \neq x \quad (2)$$

여기에서, $\widehat{U_x}$는 상품 x에 대한 추천 대상 고객 u의 선호도 예측 치이다. $\overline{U}$는 추천 대상 고객 u가 평가한 모든 상품에 대한 평균이다. J_x는 아이템 x에 대한 이웃 고객 j의 선호도 평가 치이고, $\overline{J}$는 이웃 고객 j가 평가한 모든 상품에 대한 선호도의 평균이다. $\overline{J}$의 값은 평가치 중에서 상품 x에 대한 평가치는 제외한다. r_{uj}는 추천 대상 고객 u와 추천 대상 고객의 이웃고객인 j의 선호 유사 정도를 나타내는 유사도 가중치이며, 본 논문에서는 식(1)의 피어슨 상관계수를 사용한다.

2.1 선호도 예측의 정확도 판정

협력적 필터링을 사용하여 예측의 정확도를 판정하기 위해서는 추천대상 고객이 평가한 평가 값과 협력적 필터링을 이용하여 계산된 선호도 예측 값과의 절대평균오차(Mean Absolute Error)를 사용한다. 다음 식은 선호도 예측의 정확도를 판정하기 위한 MAE에 대한 계산식이다.

$$MAE = \frac{1}{N}\sum_{j=1}^{N}|R_{uj} - \widehat{R_{uj}}| \qquad (3)$$

여기에서, R_{uj}는 상품 j에 대한 추천 대상 고객 u의 실제 선호도 평가 치이고, $\widehat{R_{uj}}$는 상품 j에 대한 추천 대상 고객 u의 협력적 필터링을 이용한 선호도 예측 값이다.

2.2 연구방법

실험 데이터는 미네소타대학의 GroupLens에서 제공된 MovieLens 100k 데이터를 실험에 사용하였으며 이 데이터는 영화에 관한 선호도를 나타낸 것이다. MovieLens 100k 데이터는 943명의 고객이 자신이 보았던 영화에 대한 의견을 평가한 1682편에 대한 자료 100,000개로 구성되어 있다. 사용자는 20편 이상의 영화에 대해 선호도를 표시하도록 설계되어 있으며, 관심의 정도에 따라 최소 1점에서 최대 5점까지 선호도를 평가할 수 있다. 본 논문에서는 GroupLens에서 제공된 MovieLens 100K dataset을 우선 80%의 training dataset 과 20%의 test dataset 으로 랜덤하게 분할하여, training dataset은 본 논문에 제안된 방법을 적용하기위해 사용하였고, test dataset은 제안한 방법의 검증을 위해 사용하였다.

그림1 dataset의 선호도 평가치 빈도분포

그림1은 MovieLens의 100K dataset 전체에 대한 선호도 평가치 빈도분포와 training dataset에 대한 선호도 평가치의 빈도 분포를 나타낸 것이다. dataset의 선호도 평가치에 대한 분포는 그림과 같으며, training dataset은 전체 dataset의 분포와 유사하여 실험에 적합하다고 판단된다.

사용된 dataset들에 대한 희소율을 조사한 결과 MovieLens 100K dataset은 94.95%의 희소율을 가지며 training dataset에 대한 선호도 평가치의 희소율은 93.90% 으로 나타났다. 이러한 희소율은 희소성이 있는 데이터로 생성되며 추천시스템의 예측 성능에 영향을 준다. 따라서 본 논문에서는 이러한 희소성 데이터를 선별하기 위한 조건을 제시한다. 다음 식은 training dataset에서 희소성이 있는 데이터를 분류하기 위하여 사용되는 선호도 평가 치에 대한 판별식이다(Kim, S. O., Lee, S.J., 2007).

$$T_k(j) = \sum_{i=1}^{user}\chi_k\,,\ k = \{1,2,3,4,5\} \qquad (4)$$

여기서, $T_k(j)$는 선호도를 k값으로 평가한 상품 j에 대한 고객의 모든 평가 값이다. 위의 식에 따라 희소성이 있는 데이터를 추출하기 위해 본 논문에서 사용된 조건은 다음과 같다.

$$\sum_{k=1}^{5} T_k(j) \le s \qquad (5)$$

여기서, j는 고객이 선호도를 표시한 상품을 나타내고 s는 희소성 데이터를 추출하기 위한 임계값이라 정한다. 임계값 s에 따라 희소성 데이터가 선택되며, 임계값보다 작은 데이터들의 집합을 집단1이라 하고, 임계값보다 큰 데이터들의 집합을 집단2라 정한다. 임계값이 작을수록 집단1은 집단2보다 희소성 데이터를 많이 포함하고 있다. 본 논문에서는 임계값에 따라 4개의 dataset을 이용하여 추천시스템의 성능을 향상시키는 방법을 제시하고자 한다. 임계값이 1인 집단2을 data1로, 임계값이 2인 집단2를 data2로 정하고 임계값이 3과 4인 집단2를 data3, data4로 정하여 이들 집단에 대하여 희소율을 조사하였다. data1보다는 data4의 집단에서 희소율이 더 커지며, 임계값이 1에서 4로 증가할수록 이들 dataset들에 대한 희소율은 증가함을 알 수 있다.

임계값	dataset	희소율(%)
1	data1	94.955
2	data2	94.958
3	data3	94.966
4	data4	94.976

표1 임계값에 따라 분류된 dataset들의 희소율

임계값에 따라 분류된 dataset들에 대한 빈도분포를 임계값에 따라 살펴보면 아래의 표2와 같으며, 임계값이 5이상인 경우 모든 dataset들의 분포가 동일하며, 임계값이 1인 경우 data1의 값이 가장 작으며 data4의 값이 가장 크다는 것을 알 수 있다. 이것은 data4보다 data1이 희소성 데이터를 많이 포함하고 있음을 의미한다. 그리고 임계값이 1인 경우 data1의 빈도수가 가장작고, 임계값이 2인 경우는 data2, 임계값이 3일 때는 data3 그리고 임계값이 4인 경우는 data4의 빈도수가 가장 작음을 알 수 있다. 임계값이 30인 경우 모든 dataset들의 빈도수가 가장 많음을 알 수 있다. 이것은 적어도 30개의 선호도 평가치가 있는 상품의 빈도가 가장 크다는 것을 나타내며 희소성 데이터가 적게 포함함을 알 수 있다.

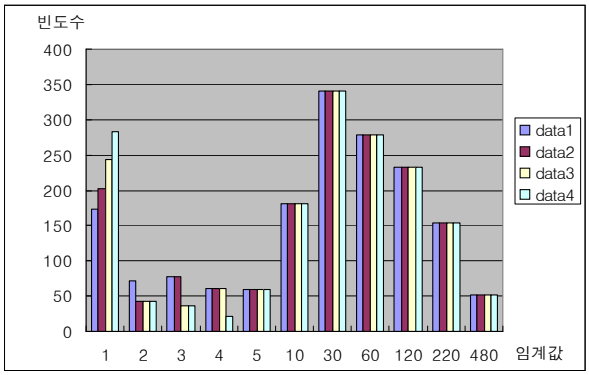

그림2 임계값에 따라 분류된 dataset들의 빈도분포

표2에서 임계값이 5미만인 경우 data4는 빈도수가 다른 dataset들보다 작으며 임계값이 4 일 때 가장 작은 빈도수를 갖는다는 것을 알 수 있다. 이는 희소성 데이터를 다른 dataset보다 적게 포함하고 있어 추천시스템의 성능을 향상시킬 가능성이 있음을 시사한다. data1의 경우는 임계값이 1일 때 다른 dataset보다는 작은 값을 갖지만 임계값이 2, 3 그리고 4일 때 다른 dataset보다 큰 값을 갖는다. 따라서 data1은 다른 dataset보다 희소성 데이터를 많이 포함한다는 것을 의미한다. 따라서 data1보다는 data4에서 임계값이 5미만인 경우 희소성 데이터가 적음을 알 수 있다.

2.3 연구결과

본 논문에서는 임계값에 따라 분류된 집단2의 집합을 4개로 나누어 data1, data2, data3 그리고 data4라 하고 이들에 대해 연구하였다. 임계값에 따라 분류된 집단1의 경우는 희소성 데이터를 많이 포함하므로 추천시스템의 예측 정확도를 떨어뜨릴 수 있다. 따라서 희소성 데이터를 덜 포함하는 집단2를 임계값에 따라 분류하고, 이 dataset들 간에 대한 예측의 정확도를 확인하기 위하여 test set을 사용하여 협력적 알고리즘을 통한 선호도 예측 값을 계산하였다. 계산된 선호도 예측의 평균은 표2와 같으며, data1의 상품에 대한 선호도 예측의 평균값은 0.7522이다. 그리고 data4의 선호도 예측의 평균값은 0.7500으로 다른 dataset들 보다 가장 작게 나타났다.

임계값	dataset	평균
1	data1	0.7522
2	data2	0.7516
3	data3	0.7510
4	data4	0.7500

표2 임계값에 따라 분류된 dataset들의 MAE 평균

임계값이 4인 data4인 경우 MAE의 평균값이 가장 작으며 이는 추천시스템에서 임계값을 4로 정하면 예측의 정확도가 임계값을 1로 정한 경우보다 좋아짐을 의미한다. 이들 집단 간의 MAE 차이를 알아보기 위하여 test dataset을 이용한 대응평균 검정 결과는 다음과 같다.

임계값	대응	N	평균	t	유의확률
1	무조건	943	0.7879	2.27	0.02*
	data1	943	0.7871		
2	무조건	943	0.7879	2.48	0.01*
	data2	943	0.7864		
3	무조건	943	0.7879	2.87	0.00**
	data3	943	0.7857		
4	무조건	943	0.7879	3.49	0.00**
	data4	943	0.7843		

*:p<0.05, **<0.01

표3 임계값에 따라 분류된 dataset들 집단 간 test dataset을 이용한 MAE의 대응평균 검정결과

data1의 MAE 평균값은 0.7871로 조건 없이 사용된 MAE 평균값 0.7879보다 좋아졌다. 그리고 data4인 경우 0.7843으로 다른 data들보다 가장 작은 값을 가진다. 따라서 본 논문에서 제시한 임계값에 따라 나누어진 집단에 대한 MAE 평균값은 기존의 방법보다는 모두 좋아졌음을 알 수 있다.

3. 결론

추천시스템에서 사용되고 있는 협력적 필터링은 고객과 이웃고객간의 선호도를 사용하여 예측 값을 생성하므로 선호도 평가치가 어느 정도는 있어야 신뢰성이 있는 추천을 할 수가 있다. 본 논문에서는 평가치가 적어 추천시스템의 신뢰를 떨어뜨리는 data들을 임계값을 사용하여 선별한 후에, 희소성 데이터를 덜 포함하는 data들을 선택하여 4개의 집단으로 나누어 이들이 추천시스템의 예측 정확도를 높일 수 있음을 실험을 통해 살펴보았다. 분석결과 조건에 따라 선별된 4개의 dataset 모두 기존의 방법보다 예측의 정확도가 높아짐이 밝혀졌다.

따라서 본 연구는 추천 시스템의 예측 정확도를 높이기 위한 하나의 방법으로 희소성이 있는 데이터를 선별하여 이들에 대한 데이터를 임계값에 따라 나누어 희소성을 완화하면 기존에 사용했던 추천시스템보다 성능이 개선됨을 알 수 있었다.

[참고문헌]

[1]김선옥, 이석준, 이희춘(2007),, "협력적 필터링에서 희소성에 따른 MAE향상에 관한 연구", 2007 한국IT서비스추계학술대회, pp.610-620, 2007.
비스학회지 제6권 제2호, 2007. 8, pp. 113~123.

[2]김재경, 오희영, 권오병(2007),, "유비쿼터스 환경에서 협업필터링을 이용한 상품그룹추천", 한국IT서비스학회지 제6권 제2호, pp. 113-123, 2007.

[3]Pazzani, M.J., CA Framework for Collaborative, Content_Based and Demograghic Filtering", Artificial Intelligent Review, pp. 394-408, 1999.

[4]Kim Hyungil, Kim Juntae., "Modifying Sparse Date for Collaborative Filtering", Journal of The Korean Society of Computer Information, pp. 610-613, Vol.32, No.1, 2005.

[5]Melville. P., Mooney. R., Nagarajan. R, "Content-Boosted Collaborative Filtering for Improved Recommendations", Proceedings of the eighteenth national Conference on Artificial Intelligence, pp. 87-192, 2002.

[6]Soboroff. I., Nocholas. C., "Combining content and collaborative in text filtering", Preceedings of the IJCAI Workshop on Machine Learning in Information Filtering, pp. 86-92, 1999.

[7]Kim, S. O., Lee, S. J., "The Effect of Data Sparsity on Prediction Accuracy in Recommender System", Journal of the Korean Society for Internet Information, Vol.8, No.6, pp. 95-102, 2007.

[8] Kim, S. O., Lee, S. J. and Lee, H. C., A Study on Improvement of Prediction Accuracy by Critical value, Journal of the Korean Data Analysis Society, Vol.10, No.1(B), pp. 591-601. 2008.

[9]Resnick, P., N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl, "GroupLens: an open architecture for collaborative filtering of netnews", in Proceedings of the 1994 ACM conference on Computer supported cooperative work. ACM Press: Chapel Hill, North Carolina, United States. pp. 175-186, 1994.

[10]J. Konstan, B. Miller, D.Maltz, J. Herlocker, L. Gordon, and J. Riedl, "GroupLens: Applying Collaborative Filtering to Usenet News", Communications of the ACM, Vol.40, No.3, pp. 77-87, 1997.