협력적 필터링에서 희소성에 따른 MAE 향상에 관한 연구

김선옥*, 이석준**, 이희춘***

*한라대학교 정보통신공학부, **상지대학교 경영학부, ***상지대학교 컴퓨터데이터정보학과

A Study on Sparsity Effect about MAE in Collaborative Filtering

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

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

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

요 약

전자상거래에서 사용되고 있는 추천시스템은 사용자들의 프로파일과 이들의 정보를 바탕으로 사용자가 선호할 만한 아이템을 추천한다. 추천시스템에서 널리 사용되고 있는 협력적 필터링 방식은 사용자들 사이의 선호도 평가치를 비교하여 유사 사용자를 선택하고, 아이템에 대한 유사 사용자의 선호도 평가치를 기반으로 하여 추천하고자 하는 아이템에 대한 사용자의 선호도를 예측하는 것이다. 하지만 사용자의 선호도가 적은 데이터로 인한 희소성 문제는 추천시스템의 성능을 저해하는 요인으로 작용하고 있다. 이러한 희소성의 문제는 선호도 평가 자료에 나타난 아이템들의 총수에 비하여 사용자가 선호한 아이템의 수가 아주 적기 때문에 발생하며, 새로운 사용자의 경우에는 아이템에 대한 선호도 평가치가 없어 유사 사용자를 선택할 수가 없어 나타나며 심한 경우에는 아이템을 전혀 추천할 수 없게 된다. 이러한 추천 시스템의 희소성문제를 해결하기 위한 방법은 희소성이 높은 데이터들에 대한 희소성을 감소시키는 것이다. 따라서 본 논문에서는 아이템에 대한 희소성을 조사하여 협력적 필터링에서 희소성 아이템이 MAE에 미치는 영향을 분석하였다. 그리고 희소성 문제를 완화하여 예측 정확도를 높이기 위한 방법으로 선호도가 적은 아이템에 대해 희소성을 최소화하는 연구와 이에 따라 희소성과 MAE의 값을 개선하는 방법을 제안한다.

1. 서론

추천시스템은 아이템에 대한 정보를 제공한 고객들의 정보를 참조하여 고객이 원하는 정보를 미리 예측하여 원하는 상품을 추천하는 시스템이다. 이 시스템은 많은 정보에 의해 결정하기가 어려울 때 가장 적합한 선택을 할 수 있도록 도움을 준다. 예를 들면 원하는 정보에 대하여 이 시스템을 이용하여 근사하게 접근시키고 고객이 원하는 상품에 대한 정보를 제공하여 고객에게 필요한 상품을 제공해준다. 전자상거래에 사용되는 추천은 고객을 위해 특별한 제공자로 사용되고 있으며 추천은 주로 고객의 정보와 다른 고객의 정보에 의존하여 원하는 아이템을 원하는 고객에게 제공한다. 컴퓨터의 발달로 인해 엄청나게 늘어나는 정보의 량을 효과적으로 신뢰할 수 있게 여과하는 일은 중요하다. 따라서 특별한 구매를 하기 위한 정보 여과에 대한 기대로 새로운 알고리즘에 대한 요구가 필요하게 되었고, 이에 따라 추천시스템이 개발되게 되었다 [1]. 이 시스템의 기초가 되는 주된 기술은 내용

기반 필터링과 협력적 필터링이며, 협력적 필터링은 미리 등록된 고객의 프로파일을 이용하여 비슷한 성향을 갖는 이웃을 선택하여 그 이웃의 정보를 바탕으로 추천대상 고객에게 정보를 제공하는 시스템이다[2]. 하지만 내용기반 필터링은 추천대상 고객 자신만의 프로파일을 이용하므로 협력적 필터링보다 적용 범위가 넓지 못하다. 협력적 필터링은 추천의 범위가 내용기반 필터링보다 광범위하지만 다른 고객의 프로파일을 이용하기 때문에 평가치가 어느 정도는 있어야 한다. 적은 평가치를 이용할 경우에는 평가 자료의 희소성으로 인해 예측의 정확도에 문제가 생긴다. 이러한 희소성의 문제를 해결하기 위하여 Pazzani[3]는 희소성의 데이터를 분리하여 속성별로 데이터를 추출하여 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 평가치에 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. 그리고 Soboroff[6]는 행렬을 이용한 SVD (Singular Value Decomposition)를 계산하여 희소성의 문제에 접근하였으며 이는 계산속도 향상에는 기여하였으나 결과적으로 정확도는 크게 나아지지 않았다. Kim[7]은 희소성의 수에 따라 집단을 분리하여 희소성이 MAE에 미치는 변화를 분석하였고, 분류된 집단에 따라 MAE의 유의적인 차이가 있음을 밝혔다. 따라서 본 논문은 희소성의 문제를 해결하기 위해 희소성이 있는 데이터를 분류하여 분류된 집단에 따라 MAE의 값을 향상시키기 위한 방법에 대한 연구로 협력적 필터링에서 선호도 평가치가 많은 고객 평가치의 평균을 평가치로 대치시키고 희소성이 있는 아이템을 추가하여 추천의 정확도가 향상됨을 조사하였다.

2. 본론

2.1 협력적 필터링

추천시스템에서 사용하는 협력적 필터링은 추천 대상고객의 선호도 평가치와 더불어 다른 고객의 선호도 평가치를 사용하여 선호도를 예측하는 알고리즘이다. 먼저 선호도를 예측하기 위하여 추천 대상 고객의 이웃을 선정하여야 하는데 모든 고객 중에서 추천대상 고객의 이웃으로 선택되기 위한 다양한 방법의 연구가 진행되고 있다. 본 논문에서 사용하는 방법은 추천 대상 고객의 추천 상품에 선호도를 평가한 고객만을 선택하여 이웃으로 선정한다. 이웃으로 선택된 고객과의 유사정도를 알기 위하여 피어슨의 상관계수를 사용한다. 다음 식은 추천대상 고객과 이웃 고객의 유사정도를 나타내는 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 유사도 가중치이며 $R_{u,i}$는 추천 대상 고객 u가 평가한 i번째 아이템에 대한 선호도 평가치이고 $\overline{R_u}$는 추천 대상 고객 u가 평가한 아이템들에 대한 평균이다. 그리고 유사도 가중치를 계산하기 위해 사용되는 평가치는 추천 대상 고객 u와 이웃고객 j가 공통으로 평가한 아이템의 평가치만 사용한다. 다음은 추천 대상 고객에게 추천할 아이템에 대한 선호도 예측으로 추천 대상 고객의 평균과 추천대상 고객의 이웃들이 평가한 평가치와 이웃의 평가치의 평균값과의 차이에 대한 값과 이웃의 유사도 가중치를 계산하여 이들 값을 합산하여 계산한다. 다음 식은 협

력적 필터링을 사용하여 선호도 예측을 계산하기 위한 알고리즘이다[9].

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

여기에서, $\hat{U}_x$는 아이템 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 평가치이고, $\overline{J}$는 이웃 고객 j가 평가한 모든 상품에 대한 선호도의 평균이다. $\overline{J}$의 값은 평가치 중에서 아이템 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.

2.2 선호도 예측의 정확도 판정

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

$$MAE = \frac{1}{N} \sum_{j=1}^{N} |R_{uj} - \hat{R}_{uj}| \quad (3)$$

여기에서, R_{uj}는 아이템 j에 대한 추천 대상 고객 u의 실제 선호도 평가 치이고, $\hat{R}_{uj}$는 아이템 j에 대한 추천 대상 고객 u의 협력적 필터링을 이용한 선호도 평가 치에 대한 예측 값이다.

2.3 연구방법

본 논문의 실험 데이터는 GroupLens에서 제공되는 MovieLens 100k 데이터를 사용하여 실험하였다. MovieLens 100k 데이터는 943명의 사용자와 이들이 선호도를 평가한 1682편의 영화에 대한 평가 자료 100,000개로 구성되어 있다. 사용자는 20편 이상의 영화에 대해 선호도를 표시하도록 설계되어 있으며, 최소 1점에서 최대 5점까지 선호도를 평가할 수 있다. 본 논문에서는 GroupLens에서 제공되는 MovieLens 100K dataset을 80%의 training dataset 과 20%의 test dataset 으로 랜덤하게 분할하여 사용하였다. 이들 dataset의 선호도 평가치에 대한 분포를 살펴보면 아래의 그림과 같으며, training dataset과 test dataset은 전제 dataset의 분포와 유사하여 실험에 적합하다고 판단된다. 그림1은 MovieLens의 100K dataset 전체에 대한 선호도 평가치 빈도분포와 training과 test dataset에 대한 선호도 평가치의 빈도 분포를 나타낸다.

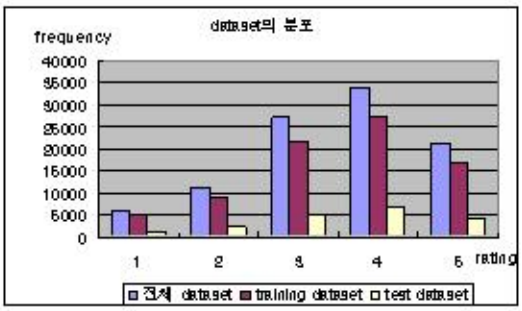

그림1 dataset의 선호도 평가치 빈도분포

희소성에 대한 영향을 연구하기위해 dataset들에 대한 희소율을 조사한 MovieLens 100K dataset과 training dataset에 대한 선호도 평가치의 희소성은 다음과 같다.

Dataset	User:Item	Sparsity(%)
Movielens 100K	943:1682	94.95
training dataset	943:1378	93.90

표1. 실험데이터의 희소성

다음 식은 training dataset에서 희소성이 있는 데이터를 분류하기 위하여 사용되는 선호도 평가치에 대한 판별식이다.

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

여기서, $T_k(j)$는 선호도를 k값으로 평가한 아이템 j 에 대한 고객의 모든 평가 값이다. 이에 따라 희소성이 있는 데이터를 추출하기 위해 본 논문에서 사용된 식은 다음과 같다.

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

여기서, j는 고객이 선호도를 표시한 아이템을 나타내고 s는 희소성을 추출하기 위한 임계값이다. 임계값 s에 따라 희소성 데이터가 선택되며 선택된 데이터들을 집단1, 나머지 데이터들을 집단2 이라 하고 이들 집단 간의 MAE 차이를 알아보기 위하여 test dataset을 이용한 t검정 결과는 다음과 같다.

s	대응	N	MAE	유의확률
4	집단1	1328	0.7739	0.000**
	집단2	49	1.7308	
6	집단1	1300	0.7581	0.000**
	집단2	77	1.6486	
8	집단1	1275	0.7436	0.000**
	집단2	102	1.6117	
15	집단1	1229	0.7205	0.000**
	집단2	148	1.5339	

**:p<0.05

표2. 희소성에 따라 분류된 training dataset에서 집단 간 test dataset의 MAE에 대한 t검정결과

t검정결과 희소성 데이터가 있는 집단과 희소성 데이터가 적게 포함된 집단 간에는 유의적인 차이가 있으며, 희소성 데이터를 많이 포함할수록 두 집단 간 MAE 차이가 커짐을 알 수 있다. 희소성 데이터를 많이 포함하는 집단2의 경우 MAE의 평균값이 1.7308 으로 임계값이 4일 때 가장 크다. 따라서 희소성이 MAE에 영향을 줌을 알 수 있다. 그리고 희소성을 완화할수록 MAE의 값이 작아져서 임계값이 15일 때 0.7205로 희소성이 적은 집단에서 가장 작게 나타났다. 따라서 예측을 정확도를 높이기 위한 희소성을 완화시키기 위한 방법으로 본 논문에서는 식(5)에 따라 희소성이 있는 아이템을 우선 선별한다. 이렇게 선별된 아이템에 선호도 응답이 많은 고객 선호도 평가치의 평균값을 선호도 평가치로 대치시켜 예측의 정확도에 대한 변화를 비교한다. 비교 결과 모든 dataset에서 MAE가 향상됨을 밝혔다.

2.4 연구결과

본 논문은 training dataset에서 희소성이 있는 아이템을 식 (5)을 사용하여 선별하고 이 아이템에 선호도 평가치를 선호도 응답이 많은 고객 선호도 평가치의 평균으로 대치시킨 후 user base에 의한 협력적 필터링을 이용하여 test dataset에 대해 선호도 평가치를 예측하였다. 예측된 평가값에 대한 예측 정확도의 변화를 조사하기 위해 본 논문에서 제안한 방법과 기존 방법에 대한 결과로 MAE를 비교하였다. 다음 표3은 임계값을 사용하여 선정된 희소성이 있는 아이템에 고객 평가치의 평균을 대치시킨 후 MAE가 개선되었는지 분석결과이다. 분석결과 희소성을 완화시킨 모든 집단에서 MAE 가 작아짐을 알 수 있다.

dataset	대응	MAE	t	자유도	유의확률
dataset1	무조건	0.7879	2.93	942	0.003**
	49개추가	0.7868			
dataset2	무조건	0.7879	3.578	942	0.000**
	77개추가	0.7860			
dataset3	무조건	0.7879	3.629	942	0.000**
	102개추가	0.7859			
dataset4	무조건	0.7879	3.715	942	0.000**
	148개추가	0.7857			

**:p<0.01

표3. 희소성이 완화된 아이템에 대한 MAE의 대응 평균 검정 결과

3. 결론

협력적 필터링을 이용한 추천시스템은 성공적으로 사용되고 있는 추천 기법이다. 하지만 협력적 필터링은 선호도 평가치가 희소한 경우 추천 성능이 저하되며, 심한 경우 추천이 전혀 이루어지지 않을 수 있다. 따라서 본 논문에서는 협력적 필터링을 이용한 추천시스템의 문제점인 희소성 문제를 해결하기 위해 고객과 아이템의 평가치와 희소성이 있는 아이템의 평가치를 선호도 평가가 많은 고객의 평균 평가 값에 적용함으로써 데이터의 희소성을 완화하였으며, 데이터의 희소

성 완화로 MAE가 개선되었음을 알 수 있었다. 따라서 본 논문에서 제안하는 방법은 실제 데이터의 희소성이 문제가 되는 전자 상거래의 추천 시스템에 효과적으로 적용될 수 있을 것이다.

[참고문헌]

[1]이희춘, 이석준, "사용자 기반 추천시스템에서 근접이웃 알고리즘과 수정알고리즘의 예측 정확도에 관한 연구", Journal of the Korean Data Analysis Society, Vol.8, No.5, 1893-1904, 2006.

[2]김재경, 오희영, 권오병(2007),, "유비쿼터스 환경에서 협업필터링을 이용한 상품그룹추천", 한국IT서비스학회지 제6권 제2호, pp. 113-123, 2007.
비스학회지 제6권 제2호, 2007. 8, pp. 113~123.

[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 Data 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, pp187-192, 2002.

[6]Soboroff, I., Nocholas, C., "Combining content and collaborative in text filtering", Preceedings of the IJCAI Workshop on Machine Learning in Information Filtering, pp86-92, 1999.

[7]Kim SunOk, Lee SeonJun., "The Effect of Data Sparsity on Prediction Accuracy in Recommender System", Journal of the Korean Society for Internet Information, Accepted, 2007.

[8]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.

[9]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.