고객의 선호도 평가패턴을 이용한 선호도 예측 알고리즘의 성능개선 방안

이석준[a], 김선옥[b], 이희춘[c]

[a] 상지대학교 경영정보학과
220-702 강원도 원주시 우산동 660번지
Tel: +82-33-738-7632, Fax: +82-33-743-1115, E-mail:digitaldesign@sangji.ac.kr

[b] 한라대학교 정보통신공학부
220-712 강원도 원주시 한라대1길 32
Tel: +82-33-760-1582, Fax: +82-33-760-1280, E-mail: sokim@halla.ac.kr

[b] 상지대학교 컴퓨터데이터정보학과
220-702 강원도 원주시 우산동 660번지
Tel: +82-33-730-0406, Fax: +82-33-730-0405, E-mail:choolee@sangji.ac.kr

Abstract

본 연구는 협업 추천 시스템에 적용되는 상품에 대한 고객의 선호도 예측 알고리즘 중 메모리기반 협업필터링 알고리즘의 선호도 예측 특성에 대하여 연구하였다. 메모리기반의 협업필터링 알고리즘은 선호도 예측 대상 고객과 유사한 성향을 가질 것으로 예상되는 고객들의 선호도 평가를 기반으로 특정 상품에 대한 선호도 예측이 이루어진다. 일반적으로 시스템을 이용하는 고객들과 선호 성향이 다른 고객들은 선호도 예측 성과가 낮은 것으로 알려져 있으며 이들이 추천시스템의 선호도 예측 정확도를 떨어뜨리는 원인으로 알려져 있다. 본 연구에서는 고객이 상품들에 평가한 선호도 평가의 패턴이 선호도 예측 정확도와 관련성이 높음을 보여 선호도 예측 알고리즘의 개선에 기초자료를 제공하고자 한다. 고객의 선호도 평가 패턴은 과거 고객이 평가한 자료로부터 얻을 수 있는 사전정보로써 선호도 예측 알고리즘을 적용하기 이전에 이용할 수 있는 정보이다. 본 연구에서는 사전정보를 이용하여 고객의 선호도 예측 오차의 특성을 연구함으로써 이들의 선호도 예측 정확도를 개선시킬 수 있는 알고리즘의 보정방법에 대하여 연구한다. 알고리즘의 보정방법을 선호도 예측 이전에 고객의 선호도 평가 특성으로 판단하여 적용함으로써 사전정보를 이용한 선호도 예측 정확도를 향상시키기 위한 접근법은 기존의 이웃 구성의 접근법과 다른 방법을 취함으로써 알고리즘 개선의 새로운 방향을 제시할 것으로 기대된다..

Keywords:

협업필터링, 사전평가, 분류함수

1. 서론

협업필터링 기법은 전자상거래에서 거래되는 상품에 대하여 목표고객의 상품들에 대한 선호도와 유사한 성향을 가진 이웃 고객들의 상품들에 대한 선호도를 이용하여 특정 상품에 대한 목표 고객의 선호도를 예측하는 기법이다. 협업필터링 기법을 이용하여 생성된 선호도 예측치를 이용하여 특정 상품에 대한 고객의 선호도를 평가하고 특정 상품들에 대한 예측 선호도의 목록을 이용하여 Top-N 목록을 작성하여 고객에게 제시함으로써 전자상거래에서 고객이 경험하게 될 상품정보의 과부하를 방지하고 함이 그 목적이다. 또한 특정 상품에 대한 고객의 선호도를 예측하고 이를 활용함으로써 개별 고객의 성향을 반영할 수 있는 개인화 서비스를 제공할 수 있다. 그렇기 때문에 고객의 선호도를 정확하게 예측 할 수 있는 선호도 예측 알고리즘의 예측 정확도는 전자상거래 웹사이트에 대한 고객의 만족도와 충성도를 높이는데 매우 중요하다. 일반적으로 협업필터링은 다수의 단점을 지니고 있지만 선호도 예측 정확도가 우수하며 상당수의 전자상거래 업체들이 이 기법을 적용하고 있는 것으로 알려져 있다. 협업필터링 기법은 전자상거래를 이용하는 모든 고객의 선호도를 이용하여 목표 고객의 선호도를 예측하기 때문에 선호도 예측 알고리즘의 적용에 있어 필요한 정보량을 줄이기 위한 노력과 선호 성향이 유사한 고객들을 선정하여 예측력을 높이기 위한 다양한 접근법이 제시되고 있다. 본 연구에서는 메모리기반의 협업필터링 알고리즘의 선호도 예측 정확도를 개선하기 위한 방법으로 기존의 접근법과 달리 개별고객이 알고리즘 적용 이전에 이미 평가한 선호 평가자료를 이용하여 선호도 예측 이전에 개별고객의 예측 정확도를 평가하는 사전평가방법과 알고리즘에 의해 생성된 선호도 예측치의 특성을

비교함으로써 선호도 예측 정확도가 낮을 것으로 예상되는 고객의 예측 정확도를 개선하기 위한 방법의 기초를 제공하고자 한다.

2. 선행연구

협업필터링의 개념은 최초의 추천시스템인 Tapestry에서 유래되었다[5]. 협업필터링 기법은 1990년대 중반 전자상거래에서 고객의 명시적 선호도를 예측하여 상품을 추천하는 추천시스템의 선호도 예측 방법으로 인지과학, 정보검색, 예측이론의 토대에서 독립적인 분야로 자리잡았다[3]. 일반적으로 추천시스템은 내용기반 기법, 협업필터링 기법, 혼합 기법으로 구분할 수 있다[4]. 협업필터링 알고리즘은 새로운 고객-상품과의 관계를 확인하기 위하여 고객과 상품 혹은 제품 간의 상호관계를 분석한다. 협력적 필터링 접근법은 상품과 고객의 과도한 정보 수집의 필요성을 피하기 위하여 고객과 상품에 대한 전문적 지식들의 수집은 배재한다. 또한, 내용기반 기법을 적용하여 파악하기 불가능하거나 어려운 숨겨진 패턴을 발견할 수 있는 가능성을 제시할 수 있다. 협력적 필터링 접근법은 학문적으로 많이 연구되고 있으며 상업적 추천시스템의 근간을 이루고 있다[6], [7].

2.1. 협업필터링 알고리즘

협업필터링 알고리즘은 협력적 필터링 알고리즘은 확률적 방법을 이용한 모형기반(model-based)의 접근법과 메모리기반(memory-based)의 접근법으로 나눌 수 있다[3],[4]. 본 연구는 메모리기반 협업필터링 알고리즘 중 Resnick 등(1994)이 제안한 이웃기반의 협업필터링 알고리즘(Neighborhood Based Collaborative Filtering Algorithm)을 적용하여 고객의 선호도를 예측한다[8]. 다음은 식(1)은 NBCFA이다.

$$\hat{U}_x = \overline{U} + \frac{\sum_{J \in Raters} \left(J_x - \bar{J}\right) \cdot r_{uj}}{\sum_{J=Raters} \left| r_{uj} \right|}, \; where \; \bar{J} = \frac{\sum_{i=1}^{n} J_i}{n}, \; i \neq x \tag{1}$$

식(1)에서 $\hat{U}_x$ 는 특정상품 x 에 대한 목표고객 U 의 선호도 예측치이며 $\overline{U}$ 는 목표고객 U 가 평가한 선호도 평가치의 평균이다. J_x 는 특정상품에 대한 이웃고객의 선호도 평가치이며 $\bar{J}$ 는 이웃고객의 선호도 평가치 평균이다. r_{uj} 는 목표고객 U 와 이웃고객 J 의 선호도 유사정도를 평가하는 유사도 가중치로 본 연구에서는 피어슨 상관계수를 이용하였다. 알고리즘에서 이웃고객의 선호도 평가치 평균은 특정 상품에 대한 이웃고객의 선호도 평가치를 제외한 평가치들로 계산된다.

일반적으로 알고리즘의 예측 정확도는 실제 선호도 평가치와 예측 선호도 평가치의 절대 편차의 평균인 MAE를 이용하여 평가한다.

$$MAE = \frac{1}{N} \sum_{j=1}^{N} \left| R_{uj} - \hat{R}_{uj} \right| \tag{2}$$

2.2. 사전평가

이석준 · 김선옥(2007)의 연구에서 개별 고객이 평가한 선호도 평가치의 통계적 특성 중 표준편차에 따라 선호도 예측 정확도가 차이가 있음을 연구하였으며 시스템 고객 전체의 선호도 평가치 분포와 다른 형태의 분포유형을 가진 고객들의 선호도 예측 정확도가 낮음을 연구하였다[1]. 또한 이석준 등(2007)의 연구에서는 개별 고객이 평가한 선호도 평가치의 특정 발생확률에 따라 선호도 예측 정확도가 낮을 고객과 반대로 선호도 예측 정확도가 높을 고객들을 정의하는 분류함수를 정의하였다[2]. 다음 식(4)와 식(5)는 선호도 예측 정확도가 낮을 고객을 선별하기 위한 기준과 분류함수이다. 먼저 고객의 선호도 평가치 R_i 를 다음 식(3)과 같이 정의한다.

$$R_i = i, \; wherer \; i = 1,2,3,4,5 \tag{3}$$

$$\delta_{u1} = \begin{cases} 1, \; f_u(R_5) \geq f_u(R_2) \\ 0, \; elsewhere \end{cases}, \; \delta_{u2} = \begin{cases} 1, \; f_u(R_1) \geq f_u(R_4) \\ 0, \; elewhere \end{cases},$$

$$\delta_{u3} = \begin{cases} 1, \; f_u(\{R_1\} \cup \{R_5\}) \geq f_u(\{R_2\} \cup \{R_3\} \cup \{R_4\}) \\ 0, \; elsewhere \end{cases} \tag{4}$$

식(4)의 δ_{u1} 조건은 개별 고객 u 의 선호도 평가치 5의 발생빈도가 선호도 평가치 2의 발생빈도 보다 클 경우 1을 그렇지 않은 경우를 0으로 정의하였다. 동일한 방법으로 δ_{u2}, δ_{u3} 의 조건을 정의한다. 정의된 조건 $\delta_{u1}, \delta_{u2}, \delta_{u3}$ 을 이용하여 다음 식(5)와 같이 분류함수를 정의한다.

$$L(\delta_{u1}, \delta_{u2}, \delta_{u3}) = \delta_{u1} \cdot \delta_{u2} \cdot \delta_{u3} \tag{5}$$

반대로 선호도 예측 정확도가 좋을 것으로 예상되는 고객을 선별하기 위한 조건과 분류함수를 다음 식(6)과 식(7)로 정의한다.

$$\theta_{u1} = \begin{cases} 1, \; f_u(R_2) \geq f_u(R_1) \\ 0, \; elsewhere \end{cases}, \; \theta_{u2} = \begin{cases} 1, \; f_u(R_4) \geq f_u(R_5) \\ 0, \; elewhere \end{cases},$$

$$\theta_{u3} = \begin{cases} 1, & f_u(R_3) \geq f_u(\{R_2\} \cup \{R_4\}) \\ 0, & elsewhere \end{cases} \quad (6)$$

식(6)은 식(4)의 내용과 동일한 방법으로 정의된다. 식(6)에 의해 정의된 $\theta_{u1}, \theta_{u2}, \theta_{u3}$ 의 조건을 이용하여 다음 식(7)과 같이 분류함수를 정의한다.

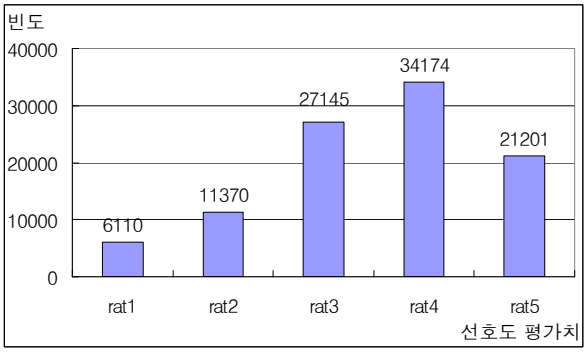
$$H(\theta_{u1}, \theta_{u2}, \theta_{u2}) = \theta_{u1} \cdot \theta_{u2} \cdot \theta_{u3} \quad (7)$$

3. 연구방법

본 연구는 기존에 제안된 사전평가의 방법에 의해 선별된 고객군의 선호도 예측 결과의 특성을 비교하여 선호도 예측 정확도를 향상시키기 위한 기초 자료를 제공하기 위하여 MovieLens 100K dataset을 이용하여 고객 선호도를 예측하였다. 일반적으로 선호도 예측 정확도의 평가는 80%의 training dataset과 20%의 test dataset으로 구분하여 20%의 test dataset에 대한 선호도 예측 정확도를 평가하지만 본 연구에서는 알고리즘의 선호도 예측 특성을 비교하기 위하여 MovieLens 100K dataset 전체 선호도 평가치에 대한 예측을 실시하였다. 선호도 예측 결과를 실제 선호도 평가치별로 구분하여 실제 선호도 평가치와 예측 선호도 평가치의 분포를 비교하였다. 또한 선호도 예측 정확도가 높을 것으로 예상되는 고객과 반대의 경우에 해당하는 고객을 선별하기 위한 분류함수를 적용하여 분류된 고객들의 선호도 예측 결과를 비교하였다.

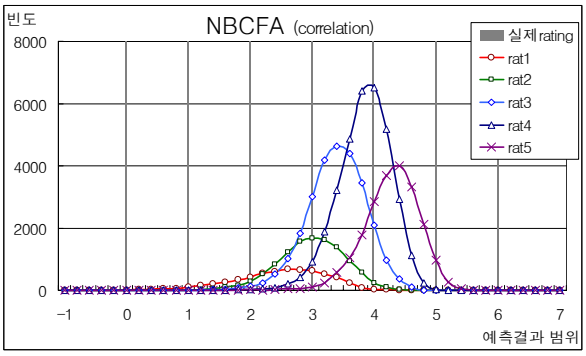
4. 실험결과

먼저 예측된 선호도 평가치와 실제 선호도 평가치의 예측 오차를 살펴보기 위하여 실제 선호도 평가치에 대한 선호도 예측치의 분포를 살펴보았다. 다음 <그림1>은 100K MovieLens dataset의 선호도 평가치 분포도로 평균은 3.53으로 우측으로 기울어져 있는 유형을 보이고 있다.



<그림1> 100K MovieLens dataset의 선호도 평가치 분포도

다음 <그림2>는 NBCFA에 의해 예측된 선호도 평가치의 분포도이다.



<그림2> 선호도 예측치의 분포도

<그림2>에서 NBCFA에 의한 예측 선호도의 분포는 모든 평가치에 대하여 정규분포의 유형을 따르고 있음을 알 수 있으며 실제 선호도 평가치 4에 대한 선호도 예측 결과의 편차가 가장 작게 나타남을 알 수 있으며 실제 선호도 평가치 1에서 3까지의 경우 높게 예측되는 것을 알 수 있다. 실제 선호도 평가치 4에서는 예측 선호도 평가치의 평균과 일치하고 있음을 알 수 있다. 실제 선호도 평가치 5에서는 반대로 낮게 선호도가 예측되는 것을 알 수 있다.

식(4)와 식(6)의 분류함수에 의해 선별된 고객 집단의 분류 정확도를 검정하기 위하여 분류집단의 개인별 MAE를 계산하고 개인별 MAE의 평균차를 분산분석을 실시하였다. 다음 <표1>은 분류된 고객 집단과 분류되지 않은 고객 집단간 분산분석 결과이다.

<표1>분산분석 결과

집단구분	N	집단 평균	표준 편차	F 값	유의 확률	사후검정
$H(\theta_{u1}, \theta_{u2}, \theta_{u2})=1$	59	0.477	0.115	53.84	0.00**	{1}{2}{3}
Non-select	869	0.576	0.161			
$L(\delta_{u1}, \delta_{u2}, \delta_{u3})=1$	15	0.952	0.126			

*: p<0.05, **: p<0.01

<표1>에서 분류함수에 의해 구분된 고객 집단간 개인별 MAE에는 차이가 있음을 알 수 있으며 사후검정 결과에서도 선별집단의 성격에 따라 잘 구분됨을 알 수 있다.

다음 <표2>는 분류함수에 희애 구분된 고객 집단에서 실제 선호도 평가치에 따라 예측된 선호도 평가치의 MAE를 나타내고 있다.

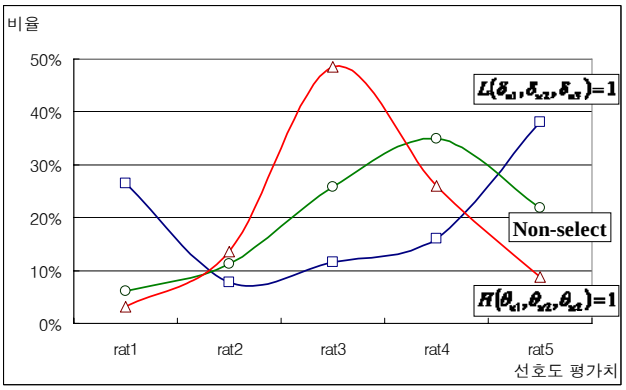
<표2>실제 선호도 평가치별 예측 선호도

	$L(\delta_{u1}, \delta_{u2}, \delta_{u3})=1$		Non-select		$H(\theta_{u1}, \theta_{u2}, \theta_{u2})=1$	
	MAE	N	MAE	N	MAE	N
rat1	1.3134	197	1.4352	5655	1.4370	200

rat2	0.9012	58	0.9335	10440	0.7427	854
rat3	0.3984	87	0.4663	23960	0.3348	3062
rat4	0.6943	120	0.3806	32385	0.4755	1646
rat5	1.1202	284	0.8237	20354	1.0568	557
total	1.0015	746	0.6264	92794	0.5251	6319

<표2>에서 실제 선호도 평가치별 예측치의 MAE는 크게 차이가 나지 않는 것을 알 수 있다. 분류함수에 따라 분류된 고객집단에서 선호도 예측 정확도가 우수할 것으로 예상되는 고객 군은 실제 선호도 평가치 2,3,4,5에서 예측 정확도가 낮을 것으로 예상되는 고객집단에 비하여 MAE가 우수하게 나타남을 알 수 있다. 분류함수에 의해 분류되지 않은 고객집단은 실제 선호도 평가치 4,5에서 분류집단의 MAE보다 우수한 것을 알 수 있다. 반면 선호도 예측 정확도가 낮을 것으로 분류된 고객집단은 실제 선호도 평가치 1에서 타 고객집단에 비하여 MAE가 우수하게 나타남을 알 수 있다.

다음 <그림2>는 실제 선호도 평가치에 따라 분류집단간 비율을 보여주고 있다.



<그림2> 선호도 예측치의 분포도

분류함수에 의해 분류된 고객집단의 비율에서 실제 선호도 평가치의 극단값인 1과 5의 비율이 높은 집단이 분류함수에 의해 분류되었으며 분류된 고객집단에서의 선호도 예측 정확도가 떨어짐을 알 수 있다. 반면 선호도 예측 정확도가 상대적으로 높은 3,4의 비율이 높은 집단이 분류함수에 의해 분류되었으며 이 집단의 선호도 예측 정확도가 타 집단에 비하여 우수함을 알 수 있다. 또한 분류함수에 의해 분류되지 않은 집단은 <그림1>의 전체 선호도 평가치 분포와 유사한 유형을 가진 고객집단임을 알 수 있다.

5. 결론

본 연구에서는 실제 선호도 평가치에 따라 NBCFA에 의해 예측된 선호도 평가치의 특성을 비교하였다. 예측 정확도를 사전에 평가하기 위한 분류함수는 개별고객이 평가한 선호도 평가의 특성으로 예측 정확도가 낮은 고객을 선별할 수 있었다. 본 연구결과를 요약하면 선호도 예측 알고리즘의 정확도는 선호도 평가치 극값에 대한 예측 정확도가 떨어지기 때문에 이를 보정해주기 위한 방법이 필요하며 이는 사용자 기반의 선호도 예측과정에 아이템에 대한 추가 정보를 이용할 필요성이 높음을 의미한다. 또한 메모리 기반의 알고리즘에 모형 기반의 알고리즘 기법을 혼용하는 하이브리드 접근법이 필요함을 의미한다. 본 연구에서는 NBCFA의 선호도 예측 특성에 대하여 연구되었지만 차기 연구로는 상품의 추가적 정보를 이용한 알고리즘의 보정과 하이브리드 기법의 적용에 관한 연구가 필요하다.

참고문헌

[1] 이석준, 김선옥 (2007). "협업필터링에서 고객의 평가치를 이용한 선호도 예측의 사전평가에 관한 연구", *경영정보학연구*, Vol.17, No.4, pp. 187-206.

[2] 이석준, 김선옥, 이희춘 (2007). "A Study on the Interrelationship between the Prediction Error and the Rating's Pattern in Collaborative Filtering", *한국데이터정보과학회지*, Vol.18, No.3, pp. 659-668.

[3] Adomavicius, G., Tuzhilin, A. (2005). "Toward the Next Generation of Recommender System: A Survey of the State-of-the-Art and Possible Extensions", *IEEE Transactions on Knowledge and Data Engineering*, Vol.17, No.6, pp. 734-749.

[4] Breese, J. S., Heckerman, D., Kadie, C. (1998). "Empirical Analysis of Predictive Algorithms for Collaborative Filtering", *In Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence*, pp. 43-52.

[5] Goldberg, D., Nichols, D., Oki, B. M. and Terry, D. (1992). "Using Collaborative Filtering to Weave an Information Tapestry", *Communications of the ACM*, Vol.35, pp. 61-70.

[6] Herlocker, J., Konstan, J., Borchers, A. and Riedl, J. (1999). "An Algorithmic Framework for Performing Collaborative Filtering", *In Proceedings of the 22nd ACM SIGIR Conference on Information Retrieval*, pp. 230-237.

[7] Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L. and Riedl, J. (1997). "GroupLens: Applying Collaborative Filtering to Usenet News", *Communications of the ACM*, Vol.40, pp. 77-87.

[8] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P. and Riedl, J. (1994). "Grouplens: An open architecture for collaborative filtering of netnews", *In Proceedings of the ACM Conference on Computer Supported Cooperative Work*, pp. 175-186.