협력적 필터링 알고리즘의 예측 성과와 사용자 선호도 평가치 특성과의 관계에 관한 연구

이 희 춘* · 이 석 준**

Abstract

본 연구는 전자상거래에서 협력적 필터링 알고리즘을 통한 사용자의 선호도 예측 정확도와 사용자가 평가한 선호도 평가치의 관계를 분석하여 알고리즘의 예측 정확도에 영향을 미치는 평가치의 통계적 특성에 관하여 연구한다. 협력적 필터링 알고리즘의 예측 정확도는 상품에 대해 공통의 관심을 갖는 이웃 사용자들의 선정과 이들의 선호도 경향이 중요한 요인이지만 본 연구에서는 선호도 예측을 위한 자신의 선호도 평가치 특성이 알고리즘에 중요한 요인임을 제시한다. 이러한 평가치의 평균, 표준편차, 왜도, 첨도 등과 같은 통계적 특성이 선호도 예측 정확도와 연관성이 있음을 제시하여 차후 연구에서 선호도 예측 이전에 사용자의 선호도 예측성과에 대한 사전평가의 가능성을 제시하고자 한다.

1. 서 론

인터넷 환경과 정보기술의 발달에 따라 추천시스템(recommender system)은 전자상거래에서 상품에 대한 고객의 정보탐색 시간을 줄이고 고객의 선호도에 부합하는 상품을 추천하여 고객의 전자상거래 이용 편의성을 높여줄 수 있는 도구로 온라인상에서 유용한 마케팅 도구로 활용되고 있다. 해외 유수의 대규모 전자상거래 사이트에서는 추천시스템을 활용하여 다년간 고객들의 서비스 만족도를 향상시키고 있다. 그렇지만 전자상거래의 규모가 확대됨에 따라 전자상거래 상에서 거래되는 상품과 고객의 수가 증가함에 따라 매우 방대한 데이터가 생성되고 있어 고객이 필요로 하는 정보만을 선택적으로 제공하는 추천시스템이 전자상거래에 필수적인 마케팅 도구로 부각되고 있다.(Kumar와 Benbasat, 2006).

* 상지대학교 컴퓨터데이터정보학과
** 상지대학교 경영정보학과

협력적 필터링(collaborative filtering) 기법은 고객 혹은 상품 간 상호 관련성을 이용하여 다른 고객 혹은 상품에 대한 선호도를 추정할 수 있어 상업적으로 가장 성공적으로 적용되고 있다. 그러나 전자상거래 상에서 거래되는 상품과 고객의 수가 증가함에 따라 상품에 대한 고객의 선호도 정보가 증가함에 따라 협력적 필터링 기법에 사용되는 데이터의 양이 매우 방대해 짐에 따라 이 선호도 정보를 모두 사용하기에는 선호도 예측에 소요되는 시간이 과도하게 소요되어 선호 정보의 적시성이 떨어지고 과도한 자원이 사용되어 효율성이 떨어질 가능성이 높아지고 있다. 이러한 변화 속에서 추천시스템을 위한 예측 알고리즘 개발과 추천품질의 향상을 위한 예측 정확도 향상에 관한 연구가 활발히 진행되고 있으며 알고리즘의 예측 특성에 대한 다양한 연구도 이루어지고 있다.(Herlocker 등, 2004; Lee 등, 2006).

본 연구에서는 협력적 필터링 기법으로 예측될 상품에 대한 고객의 선호도에 대한 사전 평가 가능성을 고객이 이미 평가한 선호도의 특성에서 찾을 수 있는지에 대해 알아보고자 함에 있다.

2. 협력적 필터링 알고리즘

협력적 필터링 기법은 전자상거래 추천 알고리즘에서 가장 핵심적인 기법으로 알려져 있으며 초기의 내용 기반의 추천시스템의 단점을 보완하고 있다. 협력적 필터링 기법은 특정 상품 혹은 아이템에 대한 목표고객의 평가치와 이웃고객의 평가치를 이용하여 목표고객이 좋아할 만한 아이템을 추천하는 기법이다(Resnick 등, 1994). 협력적 필터링 기법은 학계 및 산업계에서 널리 연구 및 적용되고 있다. 인터넷에서 협력적 필터링 기법은 Usenet 뉴스 기사에서 고객의 관심사항을 고려하여 기사 선정을 자동적으로 실행하는 연구가 진행되었고 GroupLens 연구소에서는 고객의 성향을 자동적으로 반영한 영화 추천을 위하여 MovieLens 시스템을 운용하였다.(Konstan 등, 1997). 또한 음악 추천을 위한 Ringo 시스템 등이 협력적 필터링 기법을 이용하여 적용되었다(Shardanand와 Maes, 1995). Amazon, CDNow, Netflix, MovieFinder 등에서 협업 필터링 기법을 사용하여 상품을 추천하고 있다.

협력적 필터링 기법의 가장 일반적인 알고리즘은 이웃 기반의 협력적 필터링 알고리즘(neighborhood based collaborative filtering algorithm)으로 이웃 고객들의 상품에 대한 선호 경향을 반영하여 특정 상품에 대한 추천 대상 고객의 선호도를 예측한다.

이웃 기반의 협력적 필터링 알고리즘 (NBCFA: neighborhood based collaborative filtering algorithm)은 추천 대상 고객의 평가치와 이웃으로 선정된 고객의 평가치를 이용하여 다음 식(1)과 같이 정의된다.(Resnick 등, 1994).

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

여기서, $\widehat{U_x}$는 특정 x상품에 대한 예측 대상 고객 u의 예측치로 예측 대상 고객 u가 기존에 상품들에 평가한 평가치의 평균 $\overline{U}$와 특정 상품 x에 선호도를 평가한 고객들을 이웃 고객으로 선정하고 개별 이웃 고객이 특정 상품 x에 평가한 평가치 J_x와 특정 상품 x을 제외한 평가치의 평균인 $\overline{J}$를 이용하여 선호도를 예측한다. 이때, 예측 대상 고객 u와 이웃 고객 j의 선호도 유사 정도는 유사도 가중치 r_{uj}로 정의되며 근접 이웃의 결정은 일반적으로 유사도 가중치에 의해 결정된다. 이때 유사도 가중치로는 일반적으로 상관계수가 사용된다.

선호도 예측 알고리즘의 예측 정확도는 실제 평가치와 이에 대한 예측치의 절대 오차 평균인 MAE (mean absolute error)를 이용하여 평가하며 다음 식(2)와 같이 정의한다 (Breese 등, 1998)

$$MAE = \frac{1}{N} \sum_{j=1}^{N} |R_{uj} - \widehat{R_{uj}}| \qquad 식(2)$$

3. 분석 데이터

MovieLens 100K 데이터는 GroupLens에서 943명의 사용자가 1682편의 영화에 대해 자신의 선호정도를 5점 척도로 표기한 평가치로 구성되어 있으며 개별 사용자가 최소 20편의 영화에 대해 평가한 총 10만개의 평가치로 구성되어 있다. 일반적으로 알고리즘의 예측 정확도를 평가하기 위해서는 학습 자료 (training dataset)와 평가 자료 (test dataset)로 일정 비율로 분할하여 학습 자료에 알고리즘을 적용하여 평가 자료의 선호도를 예측하는 방법을 이용한다. 그러나 본 연구에서는 충분히 데이터가 축적된 상황을 가정하여 개별 사용자가 평가한 평가치의 특성이 선호도 예측 정확도에 미치는 영향을 파악하기 위하여 10만개의 평가치 전체에 대한 예측을 실시하였다. 즉, 10만개의 평가치 중 1개의 평가치를 예측하기 위하여 99,999개의 평가치를 학습 자료로 이용하였으며 이 분석 자료를 100K_full 이라 하였다.

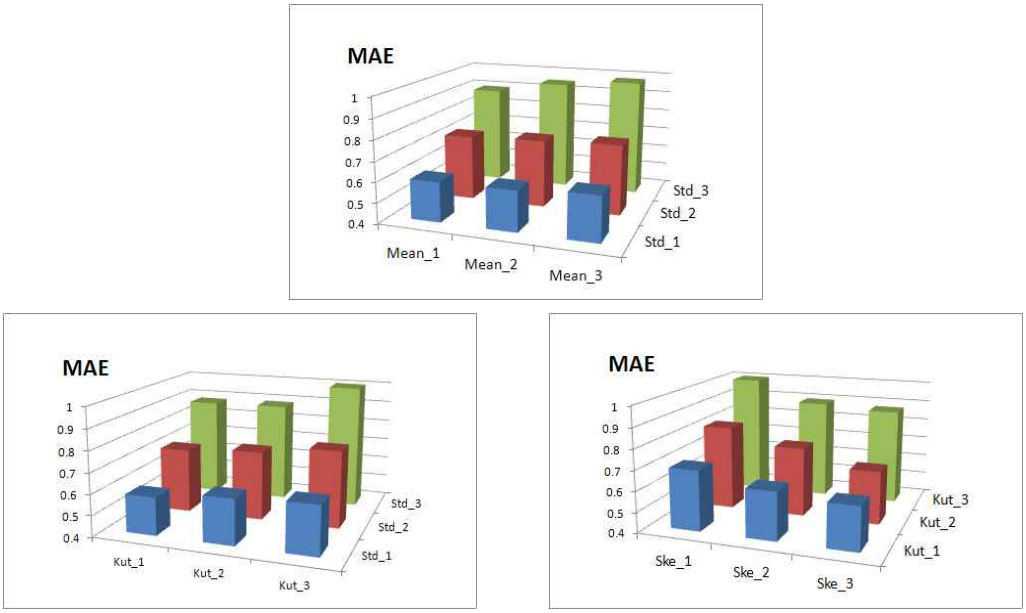
4. 분석

일반적으로 추천시스템의 정확도 평가는 전체 고객의 예측오차를 평가하는 MAE를 사용한다. 그렇지만 본 연구에서는 개별 고객이 이미 평가한 선호도 평가치가 알고리즘의 예측 정확도에 미치는 영향을 평가하기 위하여 개별 고객 즉, MovieLens 100K 데이터의 943명의 사용자에 대한 모든 평가치를 예측하여 개별 사용자별로 MAE를 계산하였다. 개별 사용자들이 평가한 선호도들의 통계적 특성과 예측 정확도인 MAE의 관계를 파악하기 위하여 개별 사용자들의 선호도 평가치의 평균, 표준편차, 첨도, 왜도를 각각 구하였다. 이들의 관계를 파악하기 위하여 먼저 각 통계량과 MAE의 상관관계 분석을 실시하였다. 각 통계량과 MAE의 상관계수는 다음 <표 1>과 같이 산정되었으며 각 통계량 중 개별 사용자의 예측 정확도인 MAE와 표준편차가 매우 높은 상관관계가 있음을 알 수 있다. 또한 평균, 첨도, 왜도와도 상관관계가 나타나고 있음을 알 수 있다. 본 연구에서는 각 통계량과 MAE의 상관관계를 이용하여 대략적인 예측 정확도를 사전에 평가할 수 있는지에 대한 가능성을 알아보기 위하여 각 통계량을 크기에 따라 3분위로 나누어 집단을 구분하고 각 집단의 MAE를 비교하였다.

<표 1> 개별 사용자의 MAE와 선호도 평가치의 통계량과의 상관분석

	MAE	평균	표준편차	첨도	왜도
MAE	1				
평균	-0.38103	1			
표준편차	0.914578	-0.40036	1		
첨도	-0.41029	0.437089	-0.36424	1	
왜도	0.147105	-0.73557	0.083824	-0.69677	1

각 통계량의 집단에 따라 예측 정확도에 차이를 알아보기 위하여 평균과 표준편차, 평균과 첨도, 첨도와 왜도의 교차집단에 따른 MAE의 교차분석 결과는 다음 [그림 1]과 같다.



[그림 1] 통계량에 의한 교차집단의 MAE 평균

[그림 1]의 상단 그림은 평균과 표준편차의 구분에 따른 9개 집단의 MAE 평균을 나타내고 있으며 하단 좌측은 첨도와 표준편차의 구분에 따른 9개 집단의 MAE 평균, 하단 우측 그림은 첨도와 왜도에 의해 구분된 집단의 MAE의 평균을 보여주고 있다.

[그림 1]을 통하여 개별 사용자의 평가치의 통계적 특성이 선호도 예측 정확도 사전 평가에 있어 유효한 구분 기준으로 사용될 수 있음을 알 수 있다.

5. 결 론

전자상거래가 활성화됨에 따라 거래되는 상품과 고객이 증가하고 있는 상황에서 고객들의 거래 정보를 이용하여 아직 구매하지 않은 상품에 대한 선호도 예측치를 생성하여 상품을 사전에 추천할 수 있는 추천시스템은 보다 고객의 구매 편의성을 높여줄 수 있는 방법으로 중요하다. 이러한 추천시스템에서 생성된 선호도 추정치에 대한 정확도를 사전에 고객에게 제공할 수 있다면 추천시스템에 대한 신뢰도를 향상시킬 수 있을 것이다. 특히 추천시스템에서 생성된 추정치를 각 개별 고객이 평가한 선호도 평가치의 기초 통계량을 이용하여 선호도 추정치의 정확도를 평가할 수 있는 방법은 추천시스템의 전반적인 정확도를 향상시키기 위해 시스템의 정확도에 부의 영향을 미칠 수 있는 고객을 선정하고 사전에 필터링하여 다양한 방법으로 시스템의 정확도를 향상시켜 줄 수 있는 방법의 개발이 가능할 것으로 생각된다. 특히 본 연구에서는 악의적 공격자의 선정과 특이 성향을 가진 고객들에 대한 조치를 사전에 고려할 수 있는 방안의 개발 가능성을 제시하고 있다.

6. 참 고 문 헌

[1] Breese, J., Heckerman, D. and Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. In Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence, 43-52.

[2] Herlocker, J., Komstan J., Terveen, L. and Riedle, J. (2004). Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems, 22, 5-53.

[3] 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, 40, 77-87.

[4] Kumar, N. and Benbasat, I. (2006). The influence of recommendation and consumer reviews on evaluations of websites, Information Systems Research, 17, 425-429.

[5] Resnick, P., Iacovou, N., Suchak, M., Bergstorm, P. and Riedl, J. (1994). GroupLens: An open architecture for collaborative filtering of netnews. In Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work, 175-186.

[6] Shardanand, U. and Maes, P. (1995). Social information filtering: Algorithms for automating 'word of mouth'. In Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems, 210-217.

[7] 이희춘, 이석준 (2006). 사용자 기반 추천시스템에서 근접이웃 알고리즘과 수정알고리즘의 예측 정확도에 관한 연구. <한국자료분석학회지>, 8, 1893-1904.