추천시스템에서 사전평가에 의해 선별된 고객의 특성에 관한 연구

임 재 화
상지대학교 경상대학 경영학과 교수
E-mail : jhim@sangji.ac.kr

이 석 준
상지대학교 경상대학 경영학과 겸임교수
E-mail : crco909@yahoo.co.kr

추천시스템은 인터넷을 기반으로 하는 전자상거래 기업에서 고객의 구매율을 높이기 위한 도구로써 이용되고 있다. 추천시스템은 전자상거래에서 거래되는 상품들에 대한 고객의 선호도를 예측하고 예측결과를 이용하여 고객들이 원하는 상품목록을 자동적으로 제시할 수 있기 때문에 고객의 정보탐색 비용을 줄여주며 동시에 고객의 구매 특성을 파악하여 마케팅 전략의 중요 자료를 제공할 수 있다. 그러나 전자상거래에서 거래되는 상품과 고객이 증가함에 따라 추천시스템은 규모의 확장성이라는 문제점을 안고 있으며 신뢰도가 낮은 추천시스템을 이용하여 고객에게 상품을 추천할 경우 추천시스템에 대한 고객의 충성도가 떨어지게 된다. 본 연구는 추천시스템에서 고객의 선호도를 예측하기 이전에 고객이 과거에 상품들에 대해 평가한 사전정보를 이용하여 예측성과에 대한 사전평가 기준을 제시하고 이를 통해 선별된 고객들의 특성에 대하여 연구하였다.

<색인어> 추천시스템, 사전평가

Ⅰ. 서 론

정보화 기술과 인터넷의 발달은 일상생활에 새로운 정보의 개념을 도입하고 있다. 기존의 상거래 역시 인터넷을 기반으로 하는 전자상거래의 도입으로 새로운 국면을 맞이하고 있으며 그 규모가 점차 확대되고 있다. 그러나 정보의 대중화는 수많은 정보를 생성하고 있으며 이는 필요한 정보만을 이용하고자 하는 사용자의 입장에서는 큰 장애물로 작용하고 있다. 전자상거래 규모는 지속적으로 증대되어 되었으며 통계청의 조사 결과에 따르면, 2006년 4/4분기 사이버 쇼핑몰 거래액은 3조 6251억원으로 3/4분기에 비해서 1711억원(5.0%)이 증가하였고 2005

논문제출일 : 2007. 7. 3, 게재확정일 : 2007. 8. 18

년 4/4분기에 비해서는 5408억원(17.5%)으로 거래액이 증가한 것으로 나타나고 있다(한국인터넷진흥원, 2007). 전자상거래 규모의 확대는 웹 상에서 거래되는 상품의 종류와 수가 증가하였음을 의미하며 이러한 상품 수의 증가에 따라 상품에 대한 정보량 또한 증대되었다. 전자상거래에서 수많은 상품 정보는 이들 상품을 구매하고자 하는 고객들에게 심각한 비용을 요구할 수 있으며 이는 고객들의 구매의욕을 저하 시킬 수 있다. 상품정보의 과잉에 따른 고객들의 구매의욕 저하를 경감시키고 개별 고객의 성향을 고려한 개인화 서비스는 전자상거래 사이트에 대한 고객들의 만족도를 높이며 이를 통한 고객의 충성도 또한 증대시킬 수 있다. 전자상거래 사이트에서 상품 정보의 과잉을 해결하고 상품에 대한 개별 고객의 선호 취향을 고려하여 고객들이 원하는 상품 정보를 제공할 수 있는 시스템이 추천시스템이다. 추천시스템은 개별 고객의 이전 상품구매 이력이나 구매 행동 패턴 등을 고려하여 고객에게 적합한 상품을 선정하여 추천하는 시스템으로 개별 고객의 선호 정보만을 이용하는 접근법과 전자상거래 시스템 내의 다른 고객들의 선호도 정보를 동시에 고려한 접근법이 있다. 상업적으로 성공적인 접근법은 고객 본인의 선호 정보뿐만 아니라 다른 고객들의 선호 정보를 동시에 고려하는 접근법으로 이를 협력적 필터링 접근법이라 한다. 본 연구는 추천시스템에서 협력적 필터링 접근법을 적용하여 선호도를 예측하기 이전에 주어진 고객들의 사전정보를 이용하여 선호도 예측 오차가 클 것으로 예상되는 고객들을 선별하고 선별된 고객집단에서 성별과 연령에 따라 예측 오차에 차이가 있는지를 통계적으로 검정하였다.

II. 선행연구

1. 협력적 필터링

협력적 필터링 기법은 상품에 대한 고객의 선호도를 예측하기 위한 알고리즘의 유형에 따라 메모리 기반(memory-based)의 알고리즘과 모형 기반(model-based)의 알고리즘으로 분류할 수 있다. 메모리 기반의 알고리즘은 고객의 선호도를 예측하기 위하여 고객이 이전에 평가한 상품과 유사한 성향이나 취향을 가진 다른 고객들이 이전에 평가한 상품들의 선호도 평가치를 이용한다. 이와 달리 모형 기반의 알고리즘은 상품에 대한 고객의 선호도를 예측하기 위하여 기계학습(machine learning), 베이지안 네트워크(bayesian network), 군집화 모형(clustering)과 같은 통계적 접근법을 취한다(Adomavicius and Tuzhilin, 2005).

협력적 필터링 접근법도 내용기반 필터링 접근법과 마찬가지로 다음과 같은 몇 가지 제약이 따른다. 첫째, 내용기반 필터링 접근법의 제약과 같이 협력적 필터링 접근법에서도 새로운 고객에 따른 문제점이 발생한다. 좀 더 정확한 선호도 예측과 상품 추천을 위해서 협력적 필터링 접근법은 고객들이 많은 선호도 평가치를 제공하여야 하는데 이는 더 많은 선호도 평가

치를 축적할수록 시스템이 고객의 선호도를 더 잘 이해할 수 있기 때문이다. 둘째, 만약 새로운 상품이 시스템에 추가될 경우 추가된 상품에 대해서 선호도를 평가한 고객들이 없기 때문에 시스템내의 고객들에게 추천이 이루어질 수 없다. 이러한 문제점을 해결하기 위해서는 시스템 관리자 혹은 추천시스템에 대한 열성적 지지자들로 구성된 패널을 이용하여 선호도 평가치를 생성하는 방법이 도입될 수 있다. 셋째, 많은 추천시스템이 안고 있는 문제점으로 고객들이 상품에 대하여 평가한 선호도 평가치의 개수가 작아서 선호도 예측이 어려운 희소성(sparsity)의 문제가 발생할 수 있다. 데이터의 희소성은 고객들과 상품들로 이루어진 행렬 구조에서 실제 선호도가 평가된 행렬의 요소 수에서 전체 요소의 수를 나눈 값으로 계산된다. 협력적 필터링 접근법을 적용하는 추천시스템의 성공은 선호도 예측을 위하여 가용한 고객들에 달려 있다고 볼 수 있다. MovieLens dataset과 같이 잘 축적된 데이터도 희소성이 대단히 크며 1 million dataset의 경우 95.8%의 희소성을 가지고 있으며 100K dataset에서도 93.7%의 희소성을 가지고 있다. 이러한 문제점들을 극복하기 위하여 다양한 접근법이 시도되고 있으며 부족한 선호도 평가치를 보충하기 위하여 고객의 인구통계학적 정보와 특성치를 이용하거나 전자상거래 사이트에서 보이는 고객의 행위에 대한 분석, 그리고 희소성을 경감시키기 위한 차원감소(dimensionality reduction)의 기법들이 적용되고 있다(김경재 · 김병국, 2005 ; 김용수, 2006).

추천시스템의 협력적 필터링 기법은 메모리 기반(memory-based)의 알고리즘과 모형 기반(model-based)의 알고리즘으로 구분한다(Breese et al., 1998). 메모리 기반의 알고리즘은 특정 상품에 대한 추천 대상 고객의 선호도를 예측하기 위하여 시스템 내의 전체 고객들의 정보를 이용하며 이웃을 선정하여 이웃과의 선호도 유사관계를 나타내는 유사도 가중치를 이용한다. 반면 모형 기반의 알고리즘은 고객들의 선호도 평가치 패턴을 기반으로 확률적 접근법의 모형을 설정하여 추천 대상 고객의 선호도 평가치를 예측하기 위하여 추천 대상 고객을 하나 혹은 다수의 소규모 계층에 분류하여 추천 상품에 대해 각 계층에 의해 예측된 평가치를 추천 대상 고객의 선호도 예측치로 사용한다. 다음 <표 1>은 메모리 기반의 협력적 필터링 알고리즘과 모형 기반의 협력적 필터링 알고리즘에 대한 분류이다.

<표 1> 협력적 필터링 알고리즘의 분류(Adomavicius and Tuzhilin, 2005 참조)

접근법	알고리즘
memory-based	NBCFA(사용자 기반, 아이템 기반) CMA(사용자 기반, 아이템 기반) 등...
model-based	neural network, k-평균 군집화, Gibbs 샘플링, 베이지안 네트워크, 나이브 베이지안, 연관규칙, LSA(Latent semantic Analysis), SVD(Singular Value Decomposition) 등...

2. 이웃 기반의 협력적 필터링 알고리즘

GroupLens에서 최초의 자동화된 이웃 기반의 협력적 필터링 알고리즘(neighborhood based collaborative filtering algorithm)을 제시하였으며 상품에 대한 추천 대상 고객의 선호도와 해당 상품에 대한 이웃 고객들의 선호도를 이용하여 추천 상품에 대한 추천 대상 고객의 선호도를 예측하게 된다(Resnick et al., 1994). 다음은 GroupLens에서 제안한 이웃 기반의 협력적 필터링 알고리즘(NBCFA)이다.

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

여기서 $\hat{U}_x$는 추천 상품 x에 대한 추천 대상 고객 u의 선호도 예측치이며, $\bar{U}$는 추천 대상 고객 u가 평가한 모든 상품에 대한 평균이다. 이때 $\bar{U}$는 추천 대상 고객의 선호 경향을 나타낸다. J_x는 추천 상품 x에 대한 이웃 고객 j의 선호도 평가치이고, $\bar{J}$는 이웃 고객 j가 평가한 모든 상품에서 추천 상품 x에 대한 평가치를 제외한 선호도 평가치들의 평균이다. 추천 대상 고객 u와 이웃 고객 j의 상품에 대한 선호도 유사정도는 유사도 가중치 r_{uj}로 정의되며 상품들에 대하여 추천 대상 고객 u와 이웃 고객 j가 공통으로 평가한 선호도 평가치로 계산된다. 일반적으로 피어슨 상관계수와 벡터 유사도가 사용되며 본 연구에서는 예측 성과가 우수한 피어슨 상관계수를 이용하여 유사도 가중치로 정의하였다(이희춘 · 이석준, 2006).

3. 예측 정확도 평가척도

알고리즘을 통한 선호도 예측치의 정확도는 실제 선호도 평가치와의 차의 절대평균인 MAE를 이용하여 평가하며 일반적으로 시스템의 정확도를 평가하기 위하여 사용된다. MAE가 크면 전체 시스템의 예측 정확도가 낮아지는 것이고 MAE가 작으면 예측 정확도가 높아진다(Shardanand and Maes, 1995). 다음은 MAE의 계산식이다.

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

여기서, R_{uj}는 상품 j에 대한 고객 u의 실제 선호도 평가치이고 $\widehat{R}_{uj}$는 상품 j에 대한 고객 u의 예측된 선호도 평가치이다.

4. 예측 오차에 대한 사전평가

본 논문은 기존에 제시된 예측 오차의 사전평가 방법에 의해 선별된 고객들의 특성에 대한 연구이다. 이희춘 등(2007)은 고객이 평가한 선호도 평가치의 통계적 특성을 이용하여 선호도 예측 오차의 사전평가 가능성에 대하여 제시하였으며 연구결과에서 선호도 예측 오차와 개별 고객이 평가한 선호도 평가치들의 표준편차가 매우 높은 관련성이 있음을 보여주고 있다. 또한 Lee 등(2007b)의 연구에서는 개별 고객의 선호도 평가치들의 표준편차의 크기에 따라 집단을 구분하여 선호도 예측 정확도를 비교한 결과 집단 간 선호도 예측 오차의 평균에서 차이가 있음을 보여주고 있다. 또한 이석준 등(2007)의 연구에서는 고객이 평가한 선호도 평가치의 분포유형과 선호도 예측 오차와의 관계를 연구하였다.

III. 실험방법

1. 예측 오차에 대한 사전평가

본 논문은 선행연구에서 제시된 연구결과를 개별 실험을 통하여 선호도 예측 오차의 사전평가에 대한 가능성과 확장성에 대하여 제시하며 선정된 고객의 성별특성과 연령별 특성에 대한 비교 분석을 한다. 실험을 위하여 본 논문에서는 MovieLens 100K dataset을 이용하였으며 선호도 예측을 위하여 NBCFA를 적용하여 선호도 예측 오차는 MAE로 측정하였다. 각 기준에 따라 분류한 고객 집단 간의 예측 오차의 차이에 대한 검정을 실시하고 결과를 바탕을 실험 dataset인 80%의 training dataset과 20%의 test dataset으로 분할하였으며 80%의 training dataset을 이용하여 20%의 test dataset에 대한 예측 결과에 동일 기준의 효과에 대하여 살펴보았다.

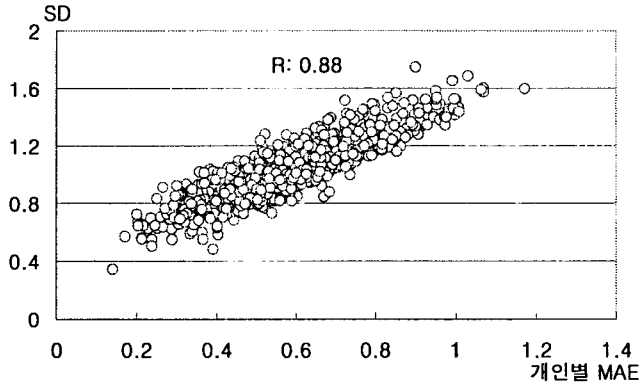
IV. 실험 및 결과

1. 표준편차를 이용한 사전평가의 분석결과

다음 <그림 1>은 MovieLens 100K dataset의 모든 선호도 평가치에 대한 선호도 예측 정

확도와 개별 고객의 표준편차와의 관계를 나타내고 있다.

<그림 1> 100K MovieLens dataset의 표준편차와 개인별 MAE와의 관계



<그림 1>을 보면 개인별 선호도 평가치의 표준편차가 크면 예측 오차가 증가하는 관계가 명확함을 알 수 있다. 이를 바탕으로 선호도 예측 이전의 개인별 사전정보인 선호도 예측치의 표준편차의 크기를 기준으로 구분한 고객 집단은 그 예측 성과에서 차이가 있을 것이다. 본 연구에서는 표준편차에 따라 구분한 집단 간 차이를 검정하고 분류된 고객 집단의 특성을 파악하기 위하여 성별과 연령에 따라 집단 내 고객들의 차이를 통계적으로 검정하였다. 다음 <표 1>은 표준편차의 크기에 따라 4분위수를 기준으로 구분한 4개의 고객 집단 간 평균차이에 대한 검정으로 100K MovieLens dataset의 전체 선호도 평가치를 이용한 결과와 80%의 training dataset을 이용한 20%의 test dataset에 대한 예측 결과를 이용한 통계검정 결과이다.

<표 2> 표준편차에 의한 구분집단 간 개인별 MAE의 기초 통계분석

구분	집단구분	N	평균	표준편차	최소값	최대값
전체 data	1집단	235	0.399	0.094	0.141	0.672
	2집단	236	0.521	0.081	0.266	0.737
	3집단	237	0.609	0.083	0.357	0.791
	4집단	235	0.775	0.121	0.510	1.171
	합계	943	0.576	0.167	0.141	1.171
실험 data	1집단	234	0.594	0.168	0.049	1.502
	2집단	237	0.697	0.169	0.354	1.567
	3집단	236	0.783	0.178	0.154	1.468
	4집단	234	1.003	0.299	0.180	2.006
	합계	941	0.769	0.259	0.049	2.006

<표 2>에서 구분집단 간 개인별 MAE의 결과에서 전체 data를 이용한 예측 결과가 더 우수함을 알 수 있으며 실험 dataset에서의 예측 수는 전체 인원에 비하여 2명의 결측 인원이 발생함을 알 수 있다. 다음 <표 3>은 구분 집단 간 분산분석 결과이다.

<표 3> 표준편차에 의해 구분된 집단 간 평균차이의 검정 결과

구분	집단구분	제곱합	자유도	평균제곱	F	유의확률	Duncan
전체 data	집단-간	17.599	3	5.866	633.636	0.000	{1}{2}{3}{4}
	집단-내	8.694	939	0.009			
	합계	26.293	942				
실험 data	집단-간	21.287	3	7.096	159.994	0.000	{1}{2}{3}{4}
	집단-내	41.556	937	0.044			
	합계	62.843	940				

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

<표 3>의 결과에서 선호도 예측 이전의 개별 고객의 선호도 평가치를 이용한 사전 평가를 통한 고객 분류는 선호도 예측 정확도를 예측 이전에 평가할 수 있는 유용한 방법임을 알 수 있다. 다음 <표 4>는 표준편차를 이용한 사전평가 방법에 의하여 구분된 고객 집단 중 선호도 예측 성능이 낮은 4집단의 고객들을 성별과 연령으로 나누어 각 집단에 속한 고객들의 인구통계변수에 따라 선호도 예측 정확도에 차이가 있는지를 검정한 결과이다.

<표 4> 성별에 따른 분류 4집단의 평균차에 대한 검정결과

구분	성별	N	평균	표준편차	평균차	t값	유의확률
전체 data	남성	156	0.770	0.117	-0.015	-0.870	0.385
	여성	79	0.784	0.130			
실험 data	남성	157	0.978	0.284	-0.077	-1.850	0.066
	여성	77	1.055	0.323			

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

<표 4>의 결과에서 표준편차에 의해 구분된 4집단의 고객들 중 성별에 따른 예측 정확도의 차이는 통계적으로 차이가 없는 것으로 나타났다.

다음 <표 5>는 고객의 연령에 따라 구분한 집단 간 평균의 차이가 있는지를 검정한 결과이다. 고객의 연령은 18세 미만, 25세 미만, 35세 미만, 45세 미만, 55세 미만, 65세 미만, 80세 미만으로 구분하여 분석하였다.

<표 5> 구분집단 내에서 연령에 따른 평균차의 검정 결과

구분	집단구분	제곱합	자유도	평균제곱	F	유의확률
전체 data	집단-간	0.111	6	0.019	1.266	0.274
	집단-내	3.333	228	0.015		
	합계	3.444	234			
실험 data	집단-간	0.436	6	0.073	0.808	0.564
	집단-내	20.402	227	0.090		
	합계	20.838	233			

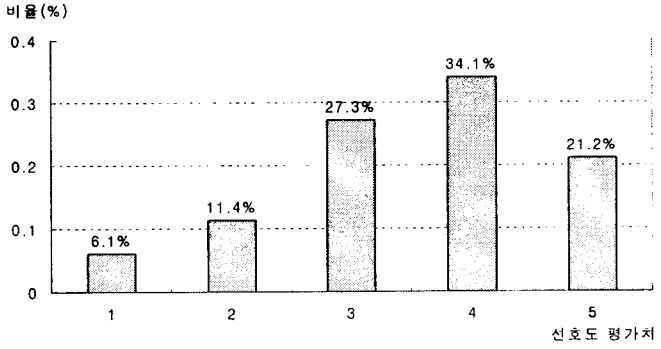
* : p<0.05,** : p<0.01

분석결과 표준편차를 이용한 분류집단에서의 성별과 연령에 따른 예측 정확도의 차이는 통계적으로 없음을 확인할 수 있으며 이는 MovieLens dataset에서 선호도 예측 성과가 낮은 고객들은 성별과 연령에 관련이 없는 것을 알 수 있다.

2. 분포 적합도 검정을 이용한 사전평가의 실험결과

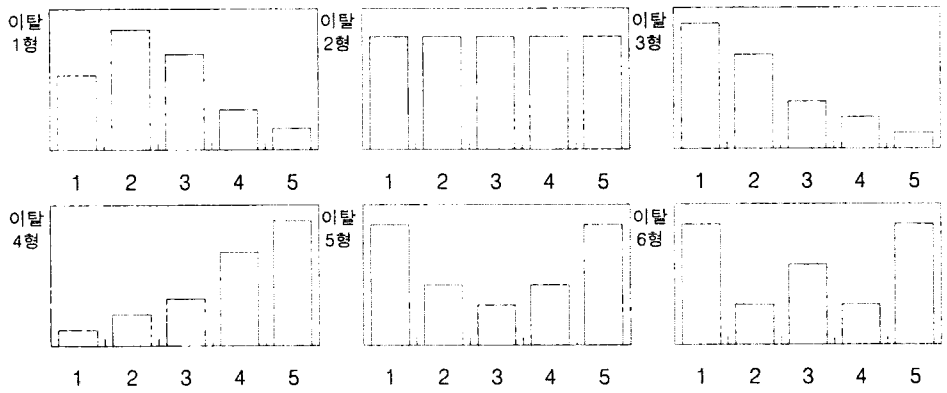
선행연구에서 개별 고객의 선호도 평가치의 분포유형에 따라 선호도 예측 정확도에 차이가 있음이 연구되었다(이석준 · 김선옥, 2007). 다음 <그림 2>는 실험 dataset에서 80%의 training dataset의 모든 선호도 평가치의 분포를 나타내고 있다. 분포의 유형은 오른쪽으로 기울어져 있는 분포의 유형을 보이고 있으며 본 연구에서는 실험 dataset의 분포 유형에 이탈하는 유형을 6가지로 가정하고 개별고객의 선호도 평가치 분포가 이탈 유형에 적합한지의 유무에 따라 구분하고 이탈 유형의 분포와 적합한 분포를 가진 고객들의 선호도 예측 성과가 낮은지를 통계적으로 검정하였다.

<그림 2> 실험 dataset에서 80%의 training dataset의 선호도 평가치 분포도



다음 <그림 3>은 <그림 2>의 분포 유형에 이탈할 것으로 가정되는 6가지의 분포 유형이다.

<그림 3> 6가지 이탈 분포 유형



<그림 3>의 분포들은 실험 dataset에서 training dataset의 분포 유형에서 이탈할 것으로 가정한 분포 형태로 training dataset의 분포와 대칭형의 분포를 이탈 제1형으로 정의하고 균등분포의 유형을 이탈 제2형, 감소형태의 분포 유형을 이탈 제3형, 증가형태의 분포 유형을 이탈 제4형, "V"형태의 분포 유형을 이탈 제5형, "W"형태의 분포 유형을 이탈 제6형으로 정의하였다. training dataset에서 정의된 이탈 유형의 분포와 적합한 선호도 평가치의 분포를 가진 고객들을 분류하기 위하여 χ^2 분포적합도 검정을 실시하여 유의수준 0.05를 기준으로 이탈 분포에 적합한 유형의 집단과 그렇지 않은 집단으로 분류하였다. 각 유형별로 집단의 개인별 MAE의 평균 차를 검정하기 위하여 독립2표본 t검정을 실시하였다. 다음 <표 6>은 χ^2 분포적합도 검정에 따라 분류된 집단 간의 독립2표본 t검정 결과이다.

<표 6> χ^2 분포적합도 검정에 따라 분류된 집단 간의 독립2표본 t검정 결과

유형	집단 구분	dataset1			
		빈도	MAE 평균	t값	유의확률
이탈 제1형	비적합	908	0.7807	-4.538	0.000**
	적합	35	0.9738		
이탈 제2형	비적합	802	0.7425	-11.177	0.000**
	적합	141	1.0461		
이탈 제3형	비적합	931	0.7860	-2.039	0.042*
	적합	12	0.9336		
이탈 제4형	비적합	700	0.7694	-3.889	0.000**
	적합	243	0.8411		
이탈 제5형	비적합	896	0.7670	-8.675	0.000**
	적합	47	1.1865		
이탈 제6형	비적합	910	0.7738	-6.416	0.000**
	적합	33	1.1764		

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

<표 6>에서 증가형태의 분포와 감소형태의 분포를 가정한 이탈 제3형과 제4형의 분석결과는 타 유형의 결과보다 상대적으로 차이가 작게 나타났다. 훈련집합의 선호도 평가치 분포형태와 대칭형의 분포 유형을 가정한 이탈 제1형도 제3형과 제4형의 분석결과보다는 평균의 차가 크게 나타났지만 균등분포의 유형으로 가정한 이탈 제2형과 "V"자 형태의 분포유형으로 가정한 이탈 제5형, "W"자 형태의 분포유형으로 가정한 이탈 제6형의 분석결과보다는 상대적으로 평균의 차가 작게 나타났다. 분석결과 훈련집합의 분포형태에서 이탈할 것으로 가정한 6개의 분포형태에 따라 분류된 고객 집단 간에는 대부분 통계적으로 유의한 차이가 있음을 알 수 있으며 이탈 제2형과 제5형, 제6형에서 분류 집단 간에 MAE의 평균의 차가 크게 나타남을 알 수 있다. 분류 집단 간 MAE의 평균의 차가 큰 각 이탈 유형에 따라 분류된 고객들의 성별과 연령에 따라 나누어 그들의 예측 정확도의 차이가 있는지를 통계적으로 검정하였다. 먼저 <표 7>은 각 이탈 유형의 적합한 고객의 집단에서 성별에 따른 선호도 예측 정확도의 차이를 알아보기 위한 t검정 결과이다.

<표 7> 각 이탈분포 유형에 따라 분류된 고객 집단에서 성별에 따른 예측 정확도의 차이에 대한 검정 결과.

구분	성별	N	평균	표준편차	평균차	t값	유의확률
이탈 제2형	남성	82	0.982	0.36538	-0.022	-0.360	0.720
	여성	58	1.004	0.35811			
이탈 제5형	남성	28	1.21	0.42462	-0.053	-0.413	0.682
	여성	18	1.263	0.43052			
이탈 제6형	남성	16	1.231	0.43288	0.051	0.352	0.727
	여성	16	1.18	0.39262			

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

<표 7>의 결과에서 이탈 분포 유형에 적합한 고객의 선정 방법은 예측 이전에 사전정보를 이용한 고객의 선호도 예측의 정확도를 평가하기에 적합한 방법이나 분류된 고객, 즉 선호도 예측 정확도가 낮을 것으로 예상된 고객의 성별에 따라서는 통계적으로 차이가 나지 않음을 알 수 있다.

다음 <표 8>은 분류고객의 연령에 따라 선호도 예측 정확도에 차이가 있는지를 검정한 결과이다.

<표 8> 각 이탈분포 유형에 따라 분류된 고객 집단에서 연령에 따른 예측 정확도의 차이에 대한 검정 결과.

구분	집단구분	제곱합	자유도	평균제곱	F	유의확률
이탈 제2형	집단-간	1.222	6	0.204	1.601	0.152
	집단-내	16.918	133	0.127		
	합계	18.141	139			
이탈 제5형	집단-간	1.047	6	0.174	0.972	0.457
	집단-내	7.003	39	0.180		
	합계	8.050	45			
이탈 제6형	집단-간	1.403	6	0.234	1.563	0.199
	집단-내	3.741	25	0.150		
	합계	5.144	31			

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

<표 8>의 결과에서도 분류된 고객들의 연령에 따라서도 통계적 차이가 발생하지 않음을 알 수 있다. 이는 선호도 예측 정확도가 낮은 고객들의 평가패턴이나 혹은 성향이 성별이나 연령에 따라 차이가 없음을 시사한다.

V. 결론

본 연구는 협력적 필터링 기법을 이용한 고객의 선호도 예측에서 예측 이전에 고객의 선호도 예측 정확도를 사전에 평가할 수 있는 선행연구의 평가방법을 적용하여 분류된 고객들이 성별과 연령에 따라 어떤 특징이 있는지를 확인하기 위하여 실험과 통계적 검정을 통하여 분석하였다. 그러나 선행연구에서 제시된 사전평가 방법들은 선호도 예측 정확도가 낮을 것으로 예상되는 고객의 분류효과는 매우 우수함을 알 수 있었지만 본 연구를 통하여 분류된 고객의 성별과 연령에 따라서는 어떠한 차이도 나지 않음을 알 수 있었다. 이는 협력적 필터링 추천시스템을 이용하는 고객의 경우 전자상거래에서 거래되는 상품에 대한 고객의 선호도가 낮은 고객들은 성별과 연령을 불문하고 어떤 공통적인 특징이 있을 것 으로 생각할 수 있으며 이를 확인하기 위해서는 추가적인 연구가 필요함을 알 수 있다. 선행연구에서 사전분류를 하지 않은 고객들의 선호도 예측 정확도는 성별과 연령에 따라서 차이가 있음이 확인되었기 때문에 선호도 예측 정확도가 낮은 고객에게서 차이가 없음은 반대로 선호도 예측 정확도가 높은 고객에게서 차이 있을 것으로 생각할 수 있다. 차기 연구로는 본 연구에서 살펴본 예측 정확도가 낮을 고객을 반대로 예측 정확도가 높을 것으로 예상되는 고객을 선정하고 이들의 특징을 살펴보는 것이 필요할 것으로 생각된다.

참 고 문 헌

김경재, 김병국(2005), "데이터 마이닝을 이용한 인터넷 쇼핑몰 상품추천시스템", 한국지능정보시스템학회논문지, 제11권, 제1호, pp. 191-205.

김용수(2006), "비정형화된 속성의 학습을 통한 자동화된 내용 기반 필터링 기법의 개발", *Journal of the Korean Data Analysis Society*, Vol.8, No.4, pp. 1615-1624.

이석준, 김선옥(2007), "협업필터링에서 고객의 평가치를 이용한 선호도 예측의 사전평가에 관한 연구", 경영정보학연구, Vol. 17, No. 4, (게제예정)

이석준, 김선옥, 이희춘(2007), "협력적 필터링에서 고객의 응답성향과 예측치의 관련성", 제9회 경영관련학회 통합학술대회, 1-8.

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

한국인터넷진흥원, "한국인터넷백서 2007", 한국인터넷진흥원, 2007.

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, 734-749.

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*, 43-52.

Lee, S. J., Kim, S. O., Lee, H. C. (2007a), "Pre-Evaluation for Detecting Abnormal Users in Recommender System", *Journal of the Korean Data & Information Society*, Vol. 18, No. 3, 619-628.

Lee, S. J., Kim, S. O., Lee, H. C. (2007b), "A Study on the Interrelationship between the Prediction Error and the Rating's Pattern in Collaborative Filtering", *Journal of the Korean Data & Information Society*, Vol. 18, No. 3, 659-668.

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* (CSCW'94), October 22-26, 175-186.

Shardanand, U. and Maes, P. (1995), "Social information filtering: algorithms for automating 'word of mouth'", *In Proceedings of the SIGCHI conference on Human factors in computing systems*, Denver, Colorado: ACM Press, 210-217.

A Study on the Features of the Classified Customers through Pre-evaluation on the Recommender System

Jae-Hwa Lim
Seok-Jun Lee

Abstract

Recommender system is the tool for E-commerce company based on the internet for increasing their sales ratio in the market. Recommender system suggests the list of items which might be wanted by customers. This list generated by the result of customers preference prediction through the prediction algorithm automatically. Recommender system will be able to offer not only the important information for marketing strategy but also reduce the cost of customers' information retrieval trough the analysis of customers' purchase patterns and features. But there are several problems like as the extension of the users and items scales and if the recommendation to customers generated by unreliable recommender system makes the customer royalty to the system to weaken. In this study, we propose the criterion for pre-evaluation on the prediction performance only using the preference ratings on the items which are rated by customers before prediction process and we study the features of customers who are classified through this classification criterion.

Key Words : Recommender System, Pre-evaluation