Journal of the Korean Data & Information Science Society
2009, **20**(3), 505–514

한국데이터정보과학회지

협력적 필터링 추천기법에서 이웃 수를 이용한 선호도 예측 정확도 향상[†]

이희춘[1]

상지대학교 컴퓨터데이터정보학과

접수 2009년 1월 22일, 수정 2009년 4월 17일, 게재확정 2009년 4월 27일

요 약

본 연구는 협력적 필터링 기법을 이용한 선호도 예측 과정에서 이웃의 수와 선호도 예측 정확도와의 관계를 분석하였다. 선호도 예측 과정에 선정된 이웃의 수를 4분위수로 4집단으로 구분하여 구분한 집단 간 선호도 예측 정확도에 차이가 나타남을 알 수 있었으며 각 집단의 예측 오차들의 평균들을 이용하여 선형의 보정함수를 제안하였다. 본 연구의 결과를 바탕으로 추천시스템에서 이웃 수를 이용한 보정함수를 이용하면 예측 정확도를 높일 수 있다.

주요용어: 전자상거래, 추천시스템, 협력적 필터링.

1. 문제제기 및 연구목적

인터넷 환경과 정보기술의 발달은 일상생활에 새로운 정보의 개념을 도입하고 있다. 기존 상거래 역시 인터넷을 기반으로 하는 전자상거래의 도입으로 새로운 국면을 맞이하고 있으며 그 규모가 점차 확대되고 있다. 전자상거래의 규모는 지속적으로 확대되었으며 한국인터넷진흥원의 자료에 따르면 오픈마켓 시장이 지속적으로 성장하고 있으며 2005년 3조원대였던 오픈마켓 시장 규모는 2006년에 5조원대로 급성장했으며 상위 7개 오픈마켓 업체의 분기 매출이 1조원대를 돌파했다. 온라인 쇼핑몰 시장도 2005년 11조원에서 2006년 13조원에 육박했다. 온라인 쇼핑몰은 소매유통업종에서 할인점, 백화점에 이어 3위의 자리를 차지할 정도로 위상이 높아졌다 (한국인터넷진흥원, 2007). 그러나 전자상거래의 규모가 확대됨에 따라 거래되는 상품과 이용 고객의 수가 증가하고 있으며 이는 곧 고객들이 기존에 상품에 대한 정보를 얻던 방식으로는 감당하기 어려운 방대한 양의 정보가 생성되고 있음을 의미한다. 그래서 대규모 상품정보에서 고객에게 필요한 정보만을 선별하여 제공하는 추천시스템이 전자상거래에 필수적인 마케팅 도구로 부각되고 있다. 또한 전자상거래의 응용 영역은 무선인터넷과 유비쿼터스 개념과 결합하여 새로운 형태로의 변화를 추구하고 있다. 이러한 변화 속에서 추천시스템을 위한 고객 선호도 예측 알고리즘 개발과 추천품질의 향상을 위한 예측 정확도 향상에 관한 연구가 활발히 진행되고 있으며 알고리즘의 선호도 예측 특성에 대한 연구도 이루어지고 있다 (김용수, 2005; Herlocker 등, 2002; Herlocker 등, 2004).

본 연구는 협력적 필터링 추천 알고리즘으로 생성된 선호도 예측치의 분석결과에서 얻어진 예측에 사용된 이웃의 수와 예측치의 관계를 이용하여 이웃 수에 따른 선호도 예측치의 보정함수를 제안하고자 한다. 이를 위하여 추천시스템의 개념과 협력적 필터링 기법의 개념, 예측 알고리즘에 대하여 살펴보고 실험 데이터 집합의 구성과 분석방법에 대하여 논하고 실험결과를 통한 보정함수의 제안으로 결론을 유도한다.

[†] 이 논문은 2007년도 상지대학교 교내 연구비 지원에 의한 것임.
[1] (220-702) 강원도 원주시 우산동 660번지, 상지대학교 컴퓨터데이터정보학과, 교수.
E-mail: choolee@sangji.ac.kr

2. 추천시스템

추천시스템은 전자상거래에서 거래되는 다양한 상품에 대한 정보 중 고객의 선호도 성향과 가장 부합할 수 있는 상품을 자동적으로 예측하여 고객에게 필터링된 정보만을 고객에게 제시할 수 있다. 이를 통하여 고객이 전자상거래 사이트에서 직접 자신의 선호도에 부합하는 상품을 찾기 위해 검색하여야 하는 많은 정보들에게 빼앗기는 시간과 비용을 줄이는 효과를 얻을 수 있으며 또한 고객들이 알지 못하던 새로운 상품 정보의 획득과 같은 효과를 얻을 수 있다. 또한 선호도 예측력이 우수한 추천시스템의 경우 고객의 특별한 선호 성향을 예측할 수 있기 때문에 개인화 서비스를 제공할 수 있다. 추천시스템은 고객에 대한 서비스 향상의 효과를 제공할 수 있을 뿐만 아니라 업체의 입장에서도 판매상품에 대한 수요예측의 자료와 목표고객의 설정과 같은 마케팅 전략에도 활용할 수 있다 (Riedl과 Konstan, 2002; Schafer 등, 2001). 추천시스템은 일반적으로 추천기법에 따라 내용 기반 (content-based), 협력적 필터링 (collaborative filtering), 혼합 필터링 (hybrid) 기법으로 구분한다.

2.1. 협력적 필터링

추천시스템에서 협력적 필터링 기법은 전자상거래 추천 알고리즘에서 가장 핵심적인 기법으로 알려져 있으며 초기의 내용 기반의 추천시스템의 단점을 보완하고 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).

협력적 필터링 기법의 가장 일반적인 알고리즘은 이웃 기반의 협력적 필터링 알고리즘으로 이웃 고객들의 상품에 대한 선호 경향을 반영하여 특정 상품에 대한 추천 대상 고객의 선호도를 예측한다. 일반적으로 이웃 기반의 협력적 필터링 알고리즘은 다음과 같은 단계로 추천 대상 고객의 특정 상품에 대한 선호도를 예측한다 (Herlocker 등, 1999).

- 1단계: 추천 대상 고객의 선호도 예측을 위한 이웃의 선정과 두 고객 간의 선호도 유사정도 측정.
- 2단계: 선호도 유사정도에 따른 예측 대상 이웃의 선정.
- 3단계: 예측 알고리즘을 이용하여 추천 대상 고객의 선호도를 예측.

이웃 기반의 협력적 필터링 알고리즘은 일반적으로 고객 간 선호도 유사정도를 이용하여 특정 상품에 대한 이웃 고객의 선호도를 예측하는 사용자 기반 (user-based)의 접근법이 적용된다. 하지만 전자상거래 사이트에서 고객의 증가는 사이트 운영자가 제어하기 어려우며 그 증가 속도가 매우 빠르기 때문에 상품 간의 유사도를 이용한 아이템 기반 (item-based)의 접근법이 적용되기도 하며 상당한 전자상거래 사이트가 아이템 기반의 기법을 적용하는 것으로 알려져 있고 (Linden 등, 2003), 또한 선호도 예측의 정확도가 높은 것으로 알려져 있다 (Sarwar 등, 2001).

2.2. 선호도 예측 알고리즘

이웃 기반의 협력적 필터링 알고리즘 (neighbor based collaborative filtering algorithm)은 추천 대상 고객의 선호도 평가치와 이웃으로 선정된 고객의 선호도 평가치를 이용하여 다음 식 (2.1)과 같이 정의된다 (Resnick 등, 1994).

$$\hat{U}_x = \bar{U} + \frac{\sum_{J \in Raters} (J_x + \bar{J}) r_{uj}}{\sum_{J \in Raters} |r_{uj}|}, \quad where\ \bar{J} = \frac{\sum_{i=1}^{n} J_i}{n},\ i \neq x. \tag{2.1}$$

여기서,

- $\hat{U}_x$: 선호도 예측 대상 상품 x 에 대한 선호도 예측 대상 고객 u 의 선호도 예 측치
- $\bar{U}$: 선호도 예측 대상 고객 u 가 평가한 모든 상품에 대한 선호도 평가치 평균
- J_x : 선호도 예측 대상 상품 x 에 대한 이웃 고객 j 의 선호도 평가치
- $\bar{J}$: 이웃 고객 j 가 평가한 모든 상품에서 선호도 예측 대상 상품 x 에 대한 평 가치를 제외한 선호도의 평균
- r_{uj} : 선호도 예측 대상 고객 u 와 이웃 고객 j 의 선호도 유사 정도를 나타내는 유사도 가중치

NBCFA는 선호도 예측에서 자신의 선호도 경향을 나타내는 $\bar{U}$ 와 이웃의 선호 경향을 나타내는 $\bar{J}$ 가 너무 과도하게 자신의 경향을 반영하기 때문에 이를 조정할 필요성이 있으며 이를 위하여 예측 대상 고객 u 와 이웃 고객 j 가 동시에 선호도를 평가한 상품만을 이용한 $\bar{U}_{match}$ 와 $\bar{J}_{match}$ 를 이용하는 대응 평균 알고리즘 (correspondence mean algorithm)이 제안되었다 (Lee, 2006).

$$\hat{U}_x = \bar{U}_{match} + \frac{\sum_{J \in Raters} (J_x - \bar{J}_{match}) r_{uj}}{\sum_{J \in Raters} |r_{uj}|}, \tag{2.2}$$

여기서,

- $\bar{U}_{match}$: 선호도 예측 대상 고객 u 와 이웃 고객 j 가 동시에 평가한 상품에 대한 선호도 평가치 평균을 모두 모아 다시 평균한 평균
- $\bar{J}_{match}$: 이웃 고객 j 가 선호도 예측 대상 고객 u 와 동시에 평가한 모든 상품에 대한 선호도의 평균

일반적으로 선호도 예측 대상 고객과 이웃고객의 선호도 유사 정도를 나타내기 위하여 다양한 형태의 유사도 가중치가 정의될 수 있으며 본 연구에서는 Pearson 상관계수를 이용한다 (Breese 등, 1998; 이희춘과 이석준, 2006). 선호도 예측 대상 고객 u 와 이웃 고객 j 의 선호도 유사 정도를 나타내는 유사도 가중치 r_{uj} 는 식 (2.3)과 같이 Pearson 상관계수로 정의한다.

$$r_{uj} = \frac{\sum_{i=1}^{m} (R_{ui} - \bar{R}_u)(R_{ji} - \bar{R}_j)}{\sqrt{\sum_{i=1}^{m} (R_{ui} - \bar{R}_u)^2 \cdot \sum_{i=1}^{m} (R_{ji} - \bar{R}_j)^2}}, -1 \le r_{uj} \le 1. \tag{2.3}$$

식 (2.3)에서 R 은 상품에 대한 고객의 선호도 평가치로 5점 척도로 되어 있으며 R_{ui} 는 상품 i 에 대한 선호도 예측 대상 고객 u 의 평가치이며 R_{ji} 는 상품 i 에 대한 고객 u 의 이웃 고객 j 의 평가치이다. $\bar{R}_u$ 와 $\bar{R}_j$ 는 고객 u 와 고객 j 가 상품들에 대한 평가치의 평균이다.

선호도 예측 알고리즘의 예측 정확도는 검증 데이터 집합의 실제 선호도 평가치와 이에 대한 선호도 예측치의 절대 오차 평균인 MAE (mean absolute error)를 이용하여 평가하며 다음 식 (2.4)와 같이 정의한다 (Breese 등, 1998; Herlocker 등, 1999; Shardanand와 Maes, 1995).

$$MAE = \frac{1}{N} \sum_{j=1}^{N} \left| R_{uj} - \hat{R}_{uj} \right|. \tag{2.4}$$

본 연구에서는 검증 데이터 집합의 개별 선호도 예측치에 오차를 줄일 수 있는 보정함수를 추가하여 생성된 새로운 예측치에 대하여 보정함수 적용 전,후의 MAE를 이용하여 선호도 예측 정확도를 비교한다.

3. 실험 설계

3.1. 실험 데이터 집합

본 연구는 GroupLens에서 공개하는 100K 와 1M MovieLens 자료를 이용하여 분석하였다. 100K MovieLens 자료는 943명의 사용자가 1682편의 영화에 대해 자신의 선호정도를 5점 척도로 표기한 선호도 평가치로 구성되어 있으며 1M 자료는 6040명의 사용자가 3952편의 영화에 대해 평가한 자료로 각각 10만개의 선호도 평가치와 1,000,209개의 선호도 평가치로 구성되어 있다. 각 사용자들은 최소 20편의 영화에 평가하도록 되어 있다. 그러나, 전체 데이터에서 50편 내의 영화를 평가한 사용자들은 100K 자료의 경우 전체의 40.3%에 해당하며 100편 내의 영화를 평가한 사용자들은 전체의 61.7%, 1M 자료의 경우 각각 29.7%, 51.8%에 해당한다. 이러한 사용자별 평가수의 불균형은 훈련과 검증 데이터 집합을 랜덤하게 분할할 경우 개별 사용자별로 균등하게 비율이 나누어지지 않는 단점이 있어 이러한 불균형을 해결하기 위하여 다음 그림 3.1과 같이 데이터를 사용자별로 분할하였다.

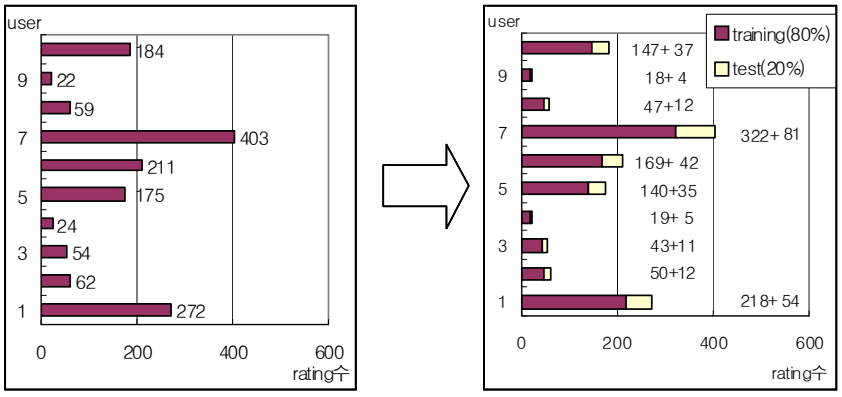

그림 3.1 사용자별 균등비율로 데이터 집합구성

최종적으로 균등비율로 분할된 사용자별 자료를 취합하여 80%의 훈련 데이터 집합과 20%의 검증 데이터 집합으로 실험에 필요한 데이터 집합을 구성하였다. 또한 선호도 예측에서 데이터의 희소성에 따라 예측 성능의 효과도 분석하기 위하여 80%의 훈련 데이터 집합에서 랜덤하게 10%씩 추출하여 구성한 70%, 60%의 훈련 데이터 집합을 구성하였다.

3.2. 이웃수의 분포

협력적 필터링 알고리즘 적용 1단계의 선호도 예측을 위한 이웃 선정 과정에서 선정된 이웃은 100K 자료와 1M 자료으로 구성한 희소성을 고려한 각각의 6:2, 7:2, 8:2 실험 데이터 집합에서 하나의 선호도 예측을 위하여 최소 1명에서 최대 466명과 2699명의 이웃이 선정되었다. 또한 실험 데이터 집합의 규모가 커질수록 예측에 선정된 이웃이 증가하고 있음을 알 수 있다. 특히 1M 자료의 실험 데이터 집합에서는 하나의 선호도 예측을 위하여 최대 2699명의 이웃의 정보가 활용되고 있음을 알 수 있다. 이러한 이웃 정보의 활용 수주의 차이가 선호도 예측 정확도에 중요한 요인이 될 가능성이 있음을 알 수 있으며 그 분포는 다음 그림 3.2와 같다.

본 연구에서는 그림 3.2의 분포도에 따라 이웃 수가 선호도 예측 정확도에 영향을 미칠 것이라 예상하

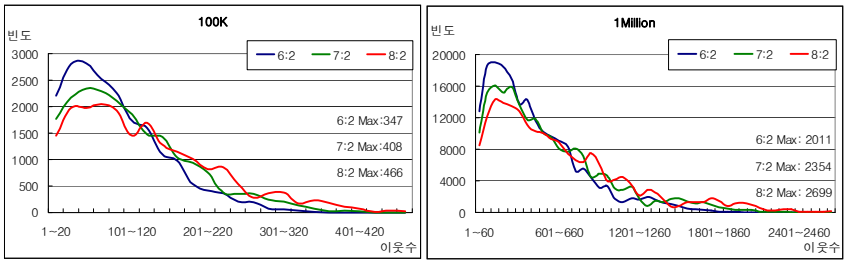

그림 3.2 100K, 1 million 실험 데이터 집합의 이웃 수 분포

고 먼저 선호도 예측 과정에 투입되는 이웃 정보의 수에 따라 선호도 예측 정확도의 차이가 있는지를 검정하였다. 실험결과에서 이웃 수에 따른 선호도 예측 정확도의 차이를 검정하고 이를 이용한 선호도 예측 알고리즘의 오차를 개선시키기 위한 보정함수를 제안한다.

4. 실험 결과

4.1. 이웃의 수와 예측치의 관계

100K 자료과 1M 자료의 각 실험 데이터 집합에서 검증 데이터 집합에 대한 개별 예측을 위해 선정된 이웃 수에 따라 선호도 예측 정확도의 차이를 검정하기 위하여 먼저 각 선호도 예측에 이용된 이웃 수들을 4분위수를 기준으로 4개의 집단으로 구분하였다. 구분 집단 간 선호도 예측 정확도의 차이가 있는지를 검정하기 위하여 크루스칼-왈리스의 순위에 의한 일원분류 분산분석을 통하여 구분 집단 간 예측 정확도에 차이가 있는지를 검정하였다. 다음 표4.1과 4.2는 100K 자료과 1M 자료의 각 실험 데이터 집합에 대한 구분 집단 간 선호도 예측 정확도 차이에 대한 분석결과이다.

표 4.1 100K 자료에서 이웃 수에 따른 집단 간 예측 정확도 차의 검정 결과

100K	집단구분		1	2	3	4	카이제곱	자유도	근사유의확률
	N=19959		5031	4992	5018	4918	-	-	-
6:2	평균	NBCFA	10804.9	10157.9	9869.5	9068.3	232.8	3	0.000**
	순위	CMA	10737.0	10135.3	9877.3	9152.8	193.4	3	0.000**
	N=19969		5022	5072	4925	4950	-	-	-
7:2	평균	NBCFA	10689.1	10197.6	9831.9	9205.1	175.9	3	0.000**
	순위	CMA	10627.9	10175.5	9852.4	9269.5	146.9	3	0.000**
	N=19973		5029	4981	4996	4967	-	-	-
8:2	평균	NBCFA	9346.0	9791.3	10132.0	10686.3	144.1	3	0.000**
	순위	CMA	9403.2	9810.7	10110.3	10630.9	120.4	3	0.000**

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

표 4.1과 표 4.2의 결과에서 100K 자료와 1M 자료의 각 실험 데이터 집합에서 이웃 수에 의해 구분된 집단 간 예측 정확도에 차이가 있음을 알 수 있다. 또한 1M 자료에 의한 결과에서 그 차이가 큼을 알 수 있다. 또한 희소성이 큰 실험 데이터 집합에서의 차이가 희소성이 완화된 실험 데이터 집합에서의 차이보다 크게 나타남을 알 수 있다. 다음 그림 4.1은 이웃 수에 의해 구분된 4개 집단의 평균 오차를 연결한 그래프이다.

표 4.2 1M 자료에서 이웃 수에 따른 집단 간 예측 정확도 차의 검정 결과

1Million	집단구분		1	2	3	4	카이제곱	자유도	근사유의확률
	N=199933		50305	49830	49878	49920	-	-	-
6:2	평균	NBCFA	103994.9	101112.9	97434.7	90931.3	1388.3	3	0.000**
	순위	CMA	103261.1	100441.3	97384.2	92369.8	908.8	3	0.000**
	N=199955		50008	50052	49944	49951	-	-	-
7:2	평균	NBCFA	102982.4	101199.1	97604.7	91481.4	1128.4	3	0.000**
	순위	CMA	102277.5	100563.6	97646.7	92763.6	727.0	3	0.000**
	N=199963		50005	50139	49943	49876	-	-	-
8:2	평균	NBCFA	103087.1	101128.4	99899.3	92633.8	898.6	3	0.000**
	순위	CMA	102627.3	100456.3	99664.1	93997.6	559.2	3	0.000**

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

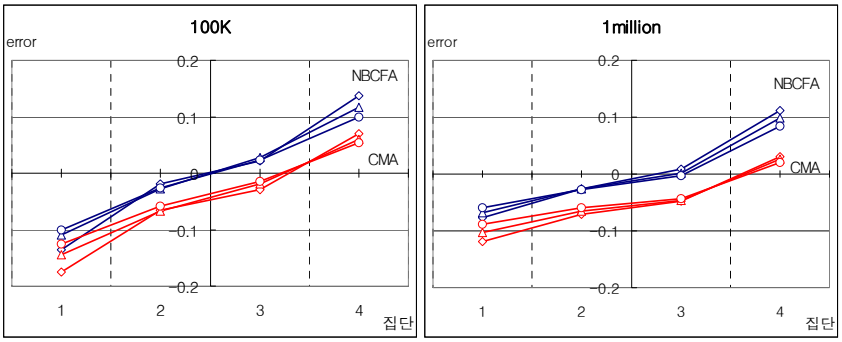

그림 4.1 100K와 1M 자료에서 구분 집단 간 오차 평균의 관계

그림에서 이웃 수가 작은 집단 1에서는 실제 선호도 평가치 보다 과대 예측되어 음의 오차가 발생하고 있으며 이웃 수가 많은 집단 4에서는 실제 선호도 평가치 보다 과소 예측되어 양의 오차가 발생하고 있음을 알 수 있다. 또한 각 집단의 오차 평균의 연결선을 통하여 이웃 수에 따라 선호도 예측 오차는 선형적인 관계를 가짐을 알 수 있다.

4.2. 이웃 수를 이용한 보정식 제안

선호도 예측 알고리즘을 이용한 예측 결과에 이웃 수를 이용한 보정함수를 추가하여 선호도 예측 정확도를 향상시키기 위하여 다음 식 (4.1)과 같은 함수를 제안한다.

$$\tilde{R}_x = \hat{R}_x + f(N_x). \tag{4.1}$$

식 (4.1)은 식 (2.1)과 (2.2)에서 제시된 NBCFA와 CMA에 의해 생성된 특정 사용자 x 의 상품들에 대한 개별 예측치에 선호도 예측 과정에서 선정된 특정 사용자 x 의 이웃 수인 N_x 를 이용한 선형보정함수 $f(N_x)$ 을 적용하여 보정된 선호도 예측치의 생성방법을 제안하고 있다. 다음 표 4.2는 이웃 수를 이용한 선호도 예측 결과의 보정을 위한 선형함수의 계수와 상수를 보여주고 있다.

다음 그림 4.2는 보정함수 적용 전 전체 선호도 평가치에 대한 MAE와 적용 후 MAE의 비교 결과이다. 실험 데이터 집합에서 희소성이 상대적으로 줄어들면 선호도 예측 정확도가 개선됨을 알 수 있으며 NBCFA에 의한 선호도 예측 결과보다 CMA에 의한 선호도 예측결과가 더 우수하게 나타나고 있음을

표 4.3 이웃 수를 이용한 보정함수

구분	실험	NBCFA		CMA	
	데이터 집합	계수	상수	계수	상수
100K	6:2	-0.00159	0.14270	-0.00142	0.17857
	7:2	-0.00113	0.12258	-0.00101	0.15336
	8:2	-0.00083	0.10731	-0.00074	0.13074
1million	6:2	-0.00023	0.09700	-0.00018	0.13090
	7:2	-0.00017	0.08769	-0.00013	0.11538
	8:2	-0.00012	0.07906	-0.00009	0.10050

알 수 있다. 또한 100K와 1M 자료 모두에서 보정함수를 적용한 선호도 예측 결과가 보정 전 예측결과보다 우수함을 알 수 있다.

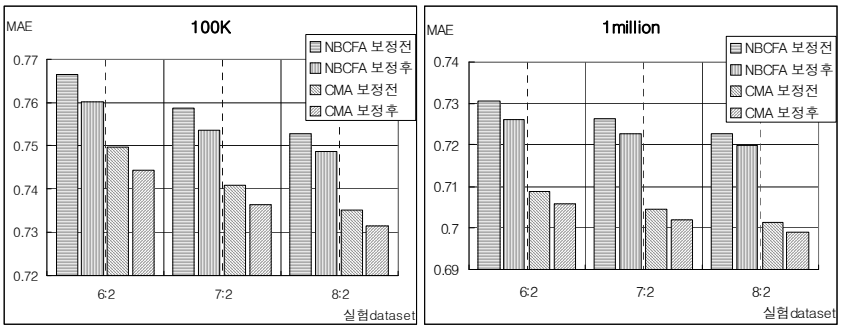

그림 4.2 보정함수 적용을 통한 선호도 예측 정확도 개선

다음 표 4.4는 보정함수 적용 전 후 예측 정확도의 차이를 검정하기 위한 대응표본 t검정 결과이다.

표 4.4 보정 전, 후 선호도 예측 정확도 차의 검정을 위한 대응표본 t검정 결과

구분	실험	NBCFA			CMA		
	데이터 집합	대응차 평균	t값	유의확률	대응차 평균	t값	유의확률
100K	6:2	0.0064	9.180	0.000**	0.0053	7.617	0.000**
	7:2	0.0049	8.176	0.000**	0.0045	7.524	0.000**
	8:2	0.0040	7.586	0.000**	0.0037	7.017	0.000**
1million	6:2	0.0045	25.572	0.000**	0.0030	17.154	0.000**
	7:2	0.0036	23.358	0.000**	0.0025	16.425	0.000**
	8:2	0.0028	20.513	0.000**	0.0020	14.962	0.000**

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

표 4.4의 결과에서 선호도 예측에 이용되는 이웃의 수를 이용하여 제안한 보정함수를 이용한 새로운 예측 결과의 예측 정확도가 보정 이전의 결과보다 통계적으로 우수한 성과를 보이고 있음을 알 수 있다.

본 연구의 결과를 통하여 데이터의 희소성이 상대적으로 감소된 8:2 실험 데이터 집합은 상대적으로 증가된 6:2 실험 데이터 집합에 비하여 선호도 예측 결과는 우수하지만 선호도 예측을 위해 선정된 이웃의 수를 이용한 보정함수를 적용하여 얻을 수 있는 예측 정확도 향상 효과가 상대적으로 작게 나타나고 있음을 알 수 있다. 본 연구에서 제안한 이웃 수를 이용한 보정함수는 100K와 1M 자료 모두에서 통계적으로 유의한 개선효과가 있음을 알 수 있으며 결과를 통하여 상품에 대한 고객의 선호도 정보가 희박

한 실제 전자상거래 데이터에 보정함수를 적용할 경우 선호도 예측 성능 향상 효과가 더 크게 나타날 수 있음을 기대할 수 있다.

5. 결론 및 시사점

본 연구는 협력적 필터링 기법을 이용한 선호도 예측 과정에서 이웃의 수와 선호도 예측 정확도와의 관계를 분석하였다. 선호도 예측 과정에 선정된 이웃의 수를 4분위수로 4집단으로 구분하여 구분한 집단 간 선호도 예측 정확도에 차이가 나타남을 알 수 있었으며 각 집단의 예측 오차들의 평균들을 이용하여 선형의 보정함수를 제안할 수 있었다. 제안한 보정함수를 통하여 100K와 1M 자료의 6:2, 7:2, 8:2 실험 데이터 집합 모두에서 보정함수를 적용하여 선호도 예측 정확도를 향상시킬 수 있었다.

그러나 본 연구에서 제한된 data를 이용한 보정함수를 제안하기 위하여 이웃 수만을 고려하여 유의한 결과를 얻었지만 data의 희소성이 상대적으로 감소된 8:2 실험 데이터 집합에서는 그 개선효과가 6:2에 비하여 작음을 알 수 있었다. 이는 좀 더 정교한 보정함수 제안을 위하여 노이즈의 제거와 새로운 변수의 추가라는 차기 연구의 필요성을 제시하고 있다. 이를 위하여 본 연구의 결과를 바탕으로 추천시스템에서 이웃 수를 이용한 보정함수를 이용하면 예측 정확도를 높일 수 있다. 본 연구에서 제안한 보정함수를 다각도로 분석하면 또 다른 보정함수를 제안할 수 있을 것으로 기대된다.

참고문헌

김용수 (2005). <전자상거래에서 고객의 탐색 및 행동 패턴을 고려한 추천시스템의 개발>, 박사학위논문, 한국과학기술원, 대전.

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

한국인터넷진흥원 (2007). <2007 한국인터넷백서>, 한국인터넷진흥원, 서울.

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.

Herlocker, J., Konstan, J., Borchers, A. and Riedl, J. (1999). An algorithm framework for performing collaborative filtering. In *Proceedings of the 1999 Conference on Research and Development in Information Retrieval,* 230-237.

Herlocker, J., Konstan, J. and Riedl, J. (2002). An empirical analysis of design choices in neighborhood based collaborative filtering algorithms. *Information Retrieval,* **5,** 287-310.

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

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.

Lee, H. C. (2006). Improved algorithm for user based recommender system. *Journal of Korean Data & Information Science Society,* **17,** 717-726.

Lee, H. C., Lee, S. J. and Chung, Y. J. (2007). A study on the improved collaborative filtering algorithm for recommender system. *SERA 2007. 5th ACIS International Conference,* 297-304.

Lee, S. J., Kim, S. O. and Lee, H. C. (2007). Pre-evaluation for detecting abnormal users in recommender system. *Journal of Korean Data & Information Science Society,* **18,** 619-628.

Linden, G., Smith, B. and York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. *Internet Computing, IEEE,* **7,** 76-80.

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.

Riedl, J. and Konstan. J. (2002). *Word of mouse: The marketing power of collaborative filtering,* Warner Books, New York.

Sarwar, B., Karypis, G., Konstan, J. and Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International World Wide Web Conference,* 285-295.

Schafer, J., Konstan, J. and Riedle, J. (2001). E-commerce recommendation applications. *Journal of Data Mining and Knowledge Discovery,* **5,** 115-152.

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.

Journal of the Korean Data & Information Science Society
2009, **20**(3), 505–514

한국데이터정보과학회지

Improving the prediction accuracy by using the number of neighbors in collaborative filtering[†]

Hee Choon Lee[1]

Department of Data Information, Sangji University

Received 22 January 2009, revised 17 April 2009, accepted 27 April 2009

Abstract

The researcher analyzes the relationship between the number of neighbors and the prediction accuracy in the preference prediction process using collaborative filtering system. The number of neighbors who are involved in the preference prediction process are divided into four groups. Each group shows a little difference in the preference prediction. By using prediction error averages in each group, linear functions are suggested. Through the result of this study, the accuracy of preference prediction can be raised when using linear functions by using the number of neighbors in the suggested system.

Keywords: Collaborative filtering, e-commerce, recommender system.

[†] This research was supported by a research fund in Sangji University (2007).
[1] Professor, Department of Data Information, Sangji University, Wonju, 220-702, Korea.
E-mail: choolee@sangji.ac.kr