Journal of the Korean Data & Information Science Society
2009, **20**(5), 809–818

한국데이터정보과학회지

근접 이웃 선정 협력적 필터링 추천시스템에서 이웃 선정 방법에 관한 연구[†]

이석준[1]

[1]상지대학교 경영정보학과

접수 2009년 7월 3일, 수정 2009년 8월 25일, 게재확정 2009년 9월 12일

요 약

협력적 필터링 기법은 전자상거래에서 거래되는 아이템에 대하여 고객들이 평가한 선호 정보를 이용하여 특정 상품에 대한 선호도 예측 대상 고객의 선호도를 예측하는 기법이다. 협력적 필터링 기법을 통한 예측 정확도를 향상시키기 위해서는 예측에 이용할 수 있는 고객들의 선호 정보를 충분히 확보하여야 한다. 그러나 과도한 이웃 고객의 선호 정보는 오히려 예측 정확도에 부정적 영향을 미치며 또한 과소 정보 역시 예측 정확도 감소에 영향을 미칠 수 있다. 본 연구에서는 협력적 필터링 알고리즘 적용에 있어 k명의 근접 이웃을 결정하는 이웃 선정방법을 개선하였으며 개별 고객의 선호도 평가 정보를 이용하여 적정 이웃 수를 결정할 수 있는 방법을 제시한다. 본 연구의 결과는 근접 이웃 수 결정을 위한 기존 방법인 탐색적 방법을 개선함과 동시에 선호도 예측 정확도를 향상 시키는데 유용한 방법을 제공할 수 있다.

주요용어: 근접 이웃 선정, 추천시스템, 협력적 필터링.

1. 문제제기 및 연구목적

추천시스템 (recommender system)은 인터넷 환경과 정보기술이 발달함에 따라 보편화된 전자상거래에서 거래 상품에 대한 고객의 정보탐색 시간을 줄일 수 있으며 웹 상에서 제공되는 질 높은 추천 서비스를 통하여 고객의 이용 편의를 증가시킬 수 있는 온라인 마케팅 도구로 활용할 수 있다. Amazon.com과 e-bay 등과 같은 대규모 전자상거래 사이트에서는 이미 추천시스템을 활용한 고객 서비스가 다년간 제공되어 왔으며 이를 통한 고객의 서비스 만족도를 향상시키고 있다. 또한 추천시스템은 전자상거래의 규모가 확대됨에 따라 거래되는 상품과 이용 고객의 수가 증가에 따른 방대한 양의 정보가 생성되고 있는 환경에서 대규모 상품정보에서 고객에게 필요한 정보만을 선별하여 제공하는 추천시스템이 전자상거래에 필수적인 마케팅 도구로 부각되고 있다 (Kumar와 Benbasat, 2006).

추천시스템의 적용을 위해서는 먼저 선호도 예측 알고리즘을 이용하여 상품에 대한 고객의 선호도를 예측하여야 하며 이를 위한 방법 중 협력적 필터링 (collaborative filtering) 기법의 알고리즘이 상업적으로 가장 성공적으로 적용되고 있다. 협력적 필터링 선호도 예측 방법은 거래 상품에 대한 고객들의 선호 정보를 이용하여 특정 상품에 대한 선호도 예측 대상 고객의 선호도를 예측하는 방법으로 예측 알고리즘을 통하여 고객의 선호도가 계산된다. 그러나 거래 상품에 대한 고객의 선호 정보가 증가함에 따라

[†] 이 논문은 2008년도 상지대학교 교내 연구비 지원에 의한 것임.
[1] (220-702) 강원도 원주시 우산동 660번지, 상지대학교 경영정보학과, 전임강사.
E-mail: digitaldesign@sangji.ac.kr

예측을 위해 사용될 수 있는 이웃 고객 정보가 증가하고 있으며 이들의 정보를 모두 사용하여 예측을 하기에는 과도한 시간이 소요된다. 이를 위해 협력적 필터링 알고리즘 적용에서 일정 수의 이웃을 선정하기 위한 k-NN (k-nearest neighbor) 방법이 적용되고 있으며 적정 이웃의 수인 k를 선정하기 위해서 유사도 가중치를 기준으로 반복적 계산 방법을 통한 탐색적 방법을 이용하여 선정하고 있다 (이재식과 박석두, 2007). 또한 전자상거래의 응용 영역은 무선인터넷과 유비쿼터스 개념과 결합하여 새로운 형태로의 변화를 추구하고 있다. 이러한 변화 속에서 추천시스템을 위한 예측 알고리즘 개발과 추천품질의 향상을 위한 예측 정확도 향상에 관한 연구가 활발히 진행되고 있으며 알고리즘의 예측 특성에 대한 연구도 이루어지고 있다 (Herlocker 등, 2004; Lee 등, 2007).

본 연구에서는 적정 이웃의 수 k의 결정을 개별 고객의 평가 정보를 이용하여 개별 고객에게 적합한 k의 결정 방법을 제시하고 이를 통한 예측 정확도를 기존의 방법을 통한 k의 선정 방법에 의한 예측 결과와 비교하여 성능을 평가한다.

2. 추천시스템

전자상거래에서 거래되는 다양한 상품에 대한 정보 중 고객의 선호도 성향과 가장 부합할 수 있는 상품을 자동적으로 예측하여 고객에게 필터링 된 정보만을 고객에게 제시할 수 있는 서비스가 개인화 서비스이다. 개인화 서비스는 인터넷 서비스 제공자들의 중요한 성공요인으로 인식되고 있다 (Ansari 등, 2000). 개인화 서비스는 개인에 대한 정보를 기반으로 서비스를 제공하기 때문에 서비스 제공자와 개인 간 정보교류가 원활할 때 효과적으로 이루어진다. 개인화 서비스 중 추천시스템은 목표고객에게 좋아할 만한 서비스나 아이템을 자동적으로 추천해주는 서비스로서 Amazon이나 CD Now 등 인터넷 쇼핑몰에서 많이 사용되고 있다. 추천시스템은 다양한 기법을 통해 구현되고 있으며 전자상거래 분야에서 쓰이는 기법 중에서 대표적인 것이 협력적 필터링 기법이다. 추천시스템은 고객이 직접 자신의 선호 성향에 부합하는 상품을 검색하는 것이 아니라 자동적으로 고객 자신의 선호 정보와 시스템 내의 이웃 정보를 이용하여 선호도를 제시하는 시스템이다. 결과적으로 추천시스템은 고객이 직접적으로 특정 상품에 대한 선호도를 요구하지 않더라도 예측치를 생성할 수 있기 때문에 이를 통한 고객에 대한 다양한 서비스 향상의 효과를 제공할 수 있을 뿐만 아니라 업체의 입장에서도 판매상품에 대한 수요예측의 자료와 목표고객의 설정과 같은 마케팅 전략에도 활용할 수 있다 (Schafer 등, 2001).

2.1. 협력적 필터링

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; Lekakos와 Giaglis, 2006). 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 알고리즘으로 이웃 고객들의 상품에 대한 선호 경향을 반영하여 특정 상품에 대한 추천 대상 고객의 선호도를 예측한다. 이웃

기반의 협력적 필터링 알고리즘을 이용한 선호도 예측은 다음과 같은 단계로 추천 대상 고객의 특정 상품에 대한 선호도 예측이 이루어진다 (이재식과 박석두, 2007).

- 1단계: 고객과 상품 간 평가치 매트릭스 구성.
- 2단계: 근접 이웃의 구성.
- 3단계: 선호도 예측 및 추천목록의 구성.

이웃 기반의 협력적 필터링 알고리즘은 고객 간 선호도 유사정도 혹은 상품 간 선호도 유사정도를 이용하여 특정 상품에 대한 이웃 고객의 선호도를 예측하며 고객 간 정보를 이용하는 방법을 사용자 기반 (user-based), 상품 간 정보를 이용하는 방법을 아이템 기반 (item-based)으로 구분한다. 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 유수의 전자상거래 사이트가 아이템 기반의 기법을 적용하는 것으로 알려져 있으며 또한 선호도 예측의 정확도가 높은 것으로 알려져 있다 (Linden 등, 2003).

2.2. 선호도 예측 알고리즘

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

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

여기서, $\widehat{U_x}$는 특정 x상품에 대한 예측 대상 고객 u의 예측치로 예측 대상 고객 u가 기존에 상품들에 평가한 평가치의 평균 $\bar{U}$와 특정 x상품에 선호도를 평가한 고객들을 이웃 고객으로 선정하고 개별 이웃 고객이 특정 상품 x에 평가한 평가치 J_x와 특정 상품 x을 제외한 평가치의 평균인 $\bar{J}$를 이용하여 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.

NBCFA는 예측치를 생성하기 위하여 예측 대상 고객 u와 그 이웃 고객들인 j의 선호 성향을 정의하기 위하여 각각이 상품들에 평가한 평가치의 평균을 이용하여 자신들의 선호 성향을 정의한다. 그러나 고객 u와 이웃 고객 j의 선호도 유사 정도를 평가하기 위한 유사도 가중치는 두 고객이 공통으로 평가한 상품에 대한 선호도 평가치 만을 이용하여 계산되기 때문에 각각의 고객이 평가한 전체의 평가치를 이용하여 선호 성향을 나타내는 것은 과도한 자신의 선호 성향을 반영하게 된다. 이러한 과도 선호 성향의 반영을 조정하기 위하여 고객 u와 이웃 고객 j의 유사도 가중치를 정의하는 것과 동일한 방법으로 평균을 구하여 사용하는 방법을 적용한 $\bar{U}_{match}$와 $\bar{J}_{match}$를 이용하는 대응평균 알고리즘 (CMA: correspondence mean algorithm)이 제안되었다 (Lee, 2006).

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

여기서, $\bar{U}_{match}$는 이웃 고객 j의 수 만큼 생성된 고객 u의 평균들을 다시 평균하여 이웃 고객들과의 관계에서 조정된 선호 성향을 나타내고 $\bar{J}_{match}$ 는 고객 u와 동시에 평가한 상품들에 대한 평가치의 평균으로 두 고객 상호간의 관계를 반영하게 된다.

2.3. 예측 성과 평가척도

선호도 예측 알고리즘의 예측 정확도는 실제 평가치와 이에 대한 예측치의 절대 오차 평균인 MAE (mean absolute error)를 이용하여 평가하며 다음 식 (2.3)과 같이 정의한다 (Breese 등, 1998; Shardanand와 Maes, 1995).

$$MAE = \frac{1}{N}\sum_{j=1}^{N}\left|R_{uj} - \widehat{R_{uj}}\right|. \tag{2.3}$$

본 연구에서는 k-NN (k-nearest neighbors) NBCFA와 CMA의 이웃 선정 방법에 따른 예측 정확도의 평가를 위하여 기존의 k의 선정 방법에 의한 예측 정확도와 예측 대상 고객의 선호 정보를 이용한 k의 선정 함수를 이용하여 개별 고객에 대한 적정 k를 선정하는 방법에 따른 선호도 예측 정확도를 비교한다. 이를 위하여 시스템 전체의 정확도는 전체 예측치의 MAE와 개별 고객의 MAE의 비교를 통하여 예측 정확도의 차이를 검정한다.

2.4. 유사도 가중치와 근접 이웃 선정

일반적으로 예측 대상 고객과 이웃고객의 선호도 유사 정도를 나타내기 위하여 다양한 형태의 유사도 가중치가 정의될 수 있으나 본 연구에서는 Pearson 상관계수를 이용한다 (Breese 등, 1998; 이희춘과 이석준, 2006). 예측 대상 고객 u와 이웃 고객 j의 선호도 유사 정도를 나타내는 유사도 가중치 r_{uj}는 식 (2.4)와 같이 Pearson 상관계수로 정의한다.

$$r_{uj} = \frac{\sum_{i=1}^{m}(R_{ui} - \bar{R}_u)(R_{ji} - \bar{R}_j)}{\sqrt{\sum_{i=1}^{m}(R_{ui} - \bar{R}_u)^2 \cdot \sum_{i=1}^{m}(R_{ji} - \bar{R}_j)^2}}, \quad -1 \le r_{uj} \le 1. \tag{2.4}$$

식 (2.4)에서 R은 상품에 대한 고객의 평가치로 5점 척도로 되어 있으며 R_{ui}는 상품 i에 대한 예측 대상 고객 u의 평가치이며 R_{ji}는 상품 i에 대한 고객 u의 이웃 고객 j의 평가치이다. $\bar{R}_u$와 $\bar{R}_j$는 고객 u와 고객 j가 상품들에 대한 평가치의 평균이다.

Pearson 상관계수와 같은 유사도 가중치는 선호도 예측 대상 고객과 선호도 예측에 필요한 이웃 고객과의 선호도 유사정도를 나타내는 척도로써 하나의 상품에 대한 선호도를 예측하기 위해서는 다수의 이웃들과의 관계가 반영된다. 이러한 다수의 유사도 가중치를 이용한 근접 이웃 고객들을 선정하는 방법에는 임계치 설정 (Shardanand and Maes, 1995) 방법과 k-NN 방법이 제안되고 있다 (Resnick 등, 1994). 임계치 설정 방법은 선호도 예측 대상 고객과 이웃 고객과의 유사도 가중치에서 미리 임계값을 설정하고 임계값 이상인 이웃들을 근접 이웃으로 선정하는 방법이고 k-NN 방법은 유사도 가중치가 높은 k명의 이웃 고객을 선정하여 이들의 정보를 선호도 예측에 반영하는 방법이다. 그러나 k명의 이웃을 선정하는 방법이 예측 대상 고객에 따라 설정되는 것이 아니라 일반적으로 전체적으로 적용되는 기존의 방법에 따라 최적의 k의 선정이 이루어진다. 다음 그림 2.1은 근접 이웃의 선정 방법을 나타내고 있으며 그림 중앙의 예측 대상 고객을 중심으로 선정 울타리가 고정되어 있으며 임계값에 의한 선정 방법이고 가변적인 울타리일 경우 k-NN 방법이라 할 수 있다.

3. 실험 설계

3.1. 실험 자료

본 연구는 GroupLens에서 공개하는 100K MovieLens 자료를 이용하여 분석하였다. 100K MovieLens 자료는 943명의 사용자가 1682편의 영화에 대해 자신의 선호정도를 5점 척도로 표기한 평가치로

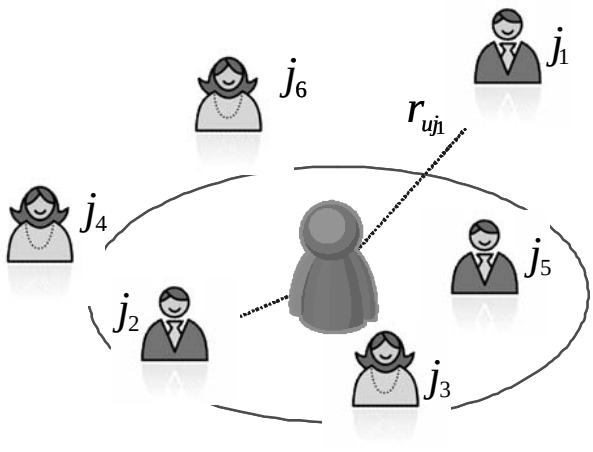

그림 2.1 근접 이웃 선정 방법

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 상황의 가정 하에 데이터가 가진 특성을 이용하여 k명의 이웃을 선정하는 방법의 가능성을 평가하기 위하여 10만개의 평가치 전체에 대한 예측을 실시하였다. 즉, 10만개의 평가치 중 1개의 평가치를 예측하기 위하여 99,999개의 평가치를 학습 자료로 이용하였으며 이 실험 자료를 100K_full 이라 하였다.

3.2. 이웃수의 분포

100K_full 실험 자료의 예측 결과 예측을 위해 생성된 이웃은 16,379,440명이다. 80%의 학습 자료와 20%의 평가 자료로 구성된 실험 자료에서 생성된 2,550,093명의 이웃과 비교하여 6.4배의 이웃이 생성되었다. 충분한 자료로 구성된 실험 자료에서 더 많은 이웃이 생성되었으며 이들을 적절히 선정하기 위해서는 근접이웃의 선정 방법이 필수적이다. 다음 그림 3.1은 100K_full 실험 자료와 8:2 실험 자료에 의해 생성된 이웃의 분포이다. 100K_full 실험 자료에서는 하나의 선호도 예측을 위하여 최대 582명의 이웃이 생성되었고 8:2 실험 자료에서는 466명의 이웃이 생성되었다.

본 연구에서는 그림 3.1의 이웃 수에서 적정 수준의 k명의 이웃을 선정하기 위한 함수를 제시하고 제안한 함수에 의해 선정된 이웃의 수를 이용한 예측 결과와 기존의 k-NN 방법으로 예측된 결과의 정확도의 차이를 검정하여 k의 선정 방법의 유효성을 제시한다.

4. 실험 결과

4.1. 기존 k-NN에 의한 근접이웃 선정

기존 연구에서는 k명의 근접이웃의 선정을 전체 예측치의 평균 MAE의 변화에 따라 탐색하고 적용하였다 (이재식과 박석두, 2007). 이는 개별 예측 대상 고객에 관계없이 일괄적으로 근접 이웃의 수가 부

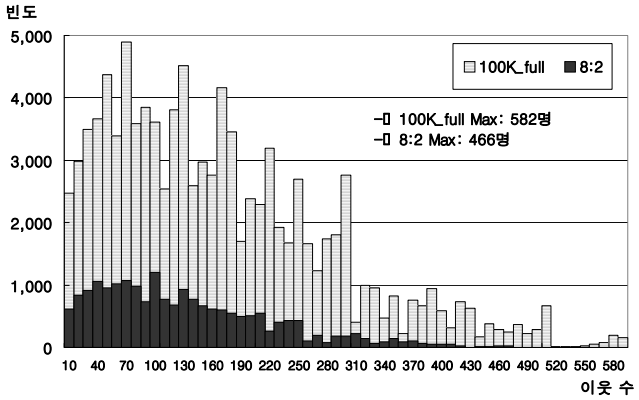

그림 3.1 100K_full 실험 자료와 8:2 실험 자료의 이웃 수 분포

여되는 방법으로 일반적으로 k-NN의 선정 방법에 이용되고 있다. 다음 그림 4.1은 기존 k-NN에 의한 근접이웃 선정 결과이다.

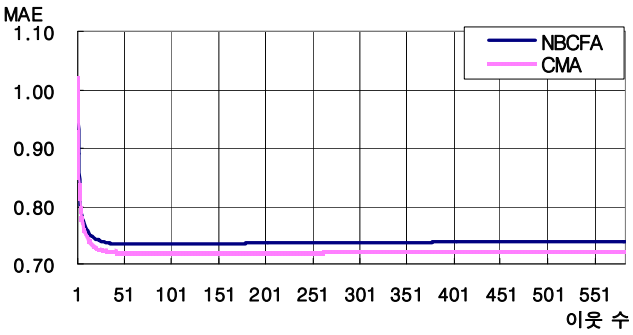

그림 4.1 기존 k-NN 방법에 의해 선정된 k에 따른 MAE 변화

그림 4.1의 결과에서 예측에 필요한 이웃 수의 변화에 따라 예측 정확도는 향상되고 있으며 이웃의 수가 50 이상부터 안정된 예측결과를 보이고 NBCFA와 CMA 모두에서 k=63의 이웃이 예측에 이용되었을 경우 가장 우수한 정확도를 나타내고 있다. 즉, 선호도 예측에 있어 63명의 이웃 이하의 경우 모든 이웃을 이용하며 63명 이상의 이웃이 있을 경우 63명의 이웃으로만 제한하여 예측하였을 경우 NBCFA에서 0.73477, CMA에서 0.71908의 결과로 예측력이 가장 우수하게 나타나고 있다. 결과에서 이웃의 수인 k가 증가할수록 MAE는 증가하여 예측력이 떨어지고 있음을 알 수 있다.

4.2. 평가치 정보를 이용한 k 예측식 제안

100K_full 실험 자료를 이용한 예측 결과를 이용하여 최적 k명의 이웃에 대한 예측의 가능성을 파악하기 위하여 실험 결과를 바탕으로 최적 k명의 이웃 수를 구하였다. 최적 k명의 이웃 수는 개별 예측 대상 고객에 따라 서로 상이한 값을 가지며 이의 예측 가능성을 평가하기 위하여 개별 고객이 평가한 예측

치의 평균, 표준편차, 각 평가치의 비율 등과 같이 예측 이전에 시스템에서 얻을 수 있는 기본 정보와 최적 이웃수와의 관계를 검토하였다. 다음 표 4.1은 고객의 평가치로부터 얻어진 정보에 따라 최적 이웃수에 차이가 있는지를 검정하기 위한 크루스칼-왈리스의 순위에 의한 일원분류 분산분석 결과이다. 결과에서 개별 고객이 평가한 평가치의 표준편차와 선호도 평가치 4, 그리고 선호도 평가치 2의 비율이 최적 이웃 수와 관계가 있는 것으로 파악되었으며 이를 검정하기 위하여 개별 고객의 평가치 표준편차, 평가치 4의 비율, 평가치 2의 비율에 따라 20, 40, 60, 80, 100분위수로 5개의 집단으로 구분하고 구분 집단 간 차이가 있는지를 검정하였다.

표 4.1 선호도 평가치의 정보에 의한 최적 이웃 수 집단 간 차이 검정 결과

집단변수	집단구분	1	2	3	4	5	카이제곱	자유도	근사유의확률
RatingSD	N=943	188	189	189	189	188	-	-	-
평균순위	633.37	551.63	429.40	373.37	372.56	136.62	4	0.000**	
R4ratio	N=943	188	189	190	190	186	-	-	-
평균순위	414.35	409.65	424.75	488.39	625.14	85.54	4	0.000**	
R2ratio	N=943	188	191	187	191	186	-	-	-
평균순위	572.09	524.01	449.12	425.37	388.32	56.83	4	0.000**	

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

표 4.1에서 개별 고객의 평가치 표준편차가 클수록 최적 이웃의 수는 적어지며 평가치 4의 비율이 높을수록 최적 이웃의 수가 커지며 평가치 2의 비율이 높을수록 최적이웃의 수가 적어지는 관계가 있음을 알 수 있다. 본 연구에서는 이들 정보를 이용한 선형 회귀식을 이용하여 최적의 이웃 수의 예측식을 다음과 같이 제안한다.

$$\widehat{y_i} = e^{(\alpha \cdot x_{i1} + \beta \cdot x_{i2} + \gamma \cdot x_{i3} + C)}. \tag{4.1}$$

식 (4.1)에서 $\widehat{y_i}$는 고객 i의 최적 근접 이웃 수에 대한 예측치이며, x_{i1}, x_{i2}, x_{i3}은 고객 i의 평가치의 표준편차, 평가치 4의 비율, 평가치 2의 비율이며, α, β, γ는 선형 회귀분석을 통하여 얻어진 각 변수의 계수이다. 본 연구에서 α는 -1.871, β는 1.065, γ는 -1.068, C는 5.799로 계산되었다.

4.3. 예측 결과의 비교

다음 그림 4.2는 최적 근접 이웃 수, 기존 k-NN 방법, 제안 예측식으로 얻어진 k명의 이웃을 이용한 선호도 예측 결과와 전체 이웃을 모두 이용한 예측 결과의 MAE 비교 결과이다.

그림 4.2의 결과에서 k-NN 방법을 적용하지 않고 모든 이웃을 이용한 예측 결과가 가장 예측 정확도가 낮으며, 기존의 k-NN 방법을 적용하여 모든 예측 대상 고객에게 동일한 k명의 이웃을 부여하는 방법이 k-NN 방법을 적용하지 않은 예측 결과보다 우수한 예측력을 보이고 있음을 알 수 있다. 본 연구에서 개별 고객의 선호도 평가 정보를 이용한 제안 방법을 이용할 경우 기존의 방법보다 향상된 예측 결과를 가짐을 알 수 있다. 결과에서 최적의 k-NN 방법을 이용한 선호도 예측 결과의 경우 가장 우수한 선호도 예측 결과를 나타내고 있으나 개별 고객에 따라 서로 상이한 최적 k명의 이웃 수를 탐색하기 위하여 모든 이웃을 이용하여 최적의 k를 탐색하여야 하는 비용이 소요된다. 그러므로 예측 대상 고객이 평가한 평가치의 정보만을 이용하여 제안 방법을 이용하면 근접 이웃의 선정에 따른 비용을 줄일 수 있다.

다음 표 4.2는 개별 방법 간 개인별 MAE에 대한 대응표본 t검정 결과이다.

표 4.2의 결과에서 k-NN 방법을 적용하지 않고 예측에서 선정된 모든 이웃을 이용하여 선호도 예측을 할 경우보다 개인별 예측 정확도에서 기존 k-NN 방법, 최적 k-NN 방법, 제안 방법 모두 통계적으로 유의한 결과를 보이고 있어 개인별 MAE에서 정확도가 향상되었음을 알 수 있다. 또한 기존의 k-NN

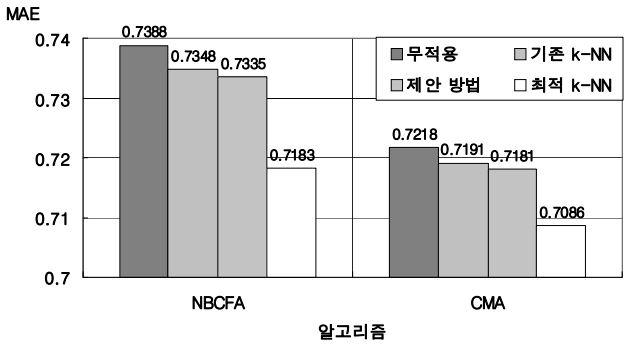

그림 4.2 개별 k 근접 이웃 선정방법에 따른 예측 결과

표 4.2 근접이웃 선정방법에 따른 개인별 MAE 비교

구분	NBCFA			CMA		
	대응차 평균	t값	유의확률	대응차 평균	t값	유의확률
무적용: 기존k-NN	0.0033	2.998	0.003**	0.0039	3.748	0.000**
무적용: 제안방법	0.0049	4.529	0.000**	0.0055	4.851	0.000**
무적용: 최적k-NN	0.0297	29.529	0.000**	0.0199	16.129	0.000**
기존k-NN: 제안방법	0.0016	3.166	0.002**	0.0015	2.705	0.007**
제안방법: 최적k-NN	0.0248	27.857	0.000**	0.0145	14.436	0.000**

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

방법과 제안 방법의 비교에서 제안 방법을 통한 k선정 결과를 이용한 개인별 MAE의 정확도가 보다 향상되었음을 알 수 있다. 그러나 최적 k-NN 결과를 이용한 선호도 예측 결과의 개인별 MAE의 정확도가 가장 높게 나타나고 있다. 결과를 통하여 제안 방법을 이용한 k-NN 방법이 기존의 방법보다 개인별 MAE의 향상에 효과가 있음을 알 수 있으며 NBCFA와 CMA 모두에서 동일한 결과를 보이고 있다.

5. 결론 및 시사점

본 연구는 협력적 필터링 기법을 이용한 선호도 예측 과정에서 적절한 근접 이웃의 수를 결정하는 방법에 대하여 연구하였으며 개별 고객이 상품에 대해 평가한 평가 정보를 이용하여 적정 k 근접 이웃의 수를 결정할 수 있는 방법에 대하여 제안하였다. 분석 결과에서 k-NN 방법을 적용하지 않았을 때 가장 예측력이 가장 낮게 나타났으며 기존의 k-NN 방법을 적용할 경우 보다 향상된 선호도 예측 결과를 얻을 수 있었으며 본 연구에서 제안한 방법을 통하여 개별 고객에 따라 예측된 k명의 이웃을 적용한 결과에서 기존의 방법보다 향상된 선호도 예측 결과를 얻을 수 있었다. 최적의 k-NN 선정을 통한 선호도 예측 정확도가 가장 우수한 성과를 나타내고 있지만 실제 최적 k명의 이웃 선정 방법은 최적 k 탐색 시간의 소요로 신속한 선호도 예측치 생성에 있어서는 적합하지 않을 것으로 예상된다. 따라서 본 연구에서 제안된 방법을 이용하여 k를 선정할 경우 최적 k의 선정에 소요되는 자원에 상응하는 효과를 거둘 수 있을 것을 기대된다.

그러나 본 연구에서는 고객이 평가한 선호도 정보가 충분한 상황을 가정하여 모든 평가치를 이용한 예측 결과를 이용하였으며 희소 데이터의 상황에서 본 연구에서 제안한 방법의 적용에 대해서는 추가적인 연구가 필요하다. 또한 개별 고객의 정보만을 이용하여 적정 k를 선정하는 방법의 정확도를 보다 높이

기 위해 상품의 정보를 이용하는 방법에 대한 추가적인 연구가 요구된다. 그리고 본 연구에서는 개별 고객에게 적합한 k의 선정 방법에 대한 연구이지만 개별 선호도 예측 결과에 있어서는 일괄적인 k를 부여하는 방법이기 때문에 개별 선호도 예측 결과에 적용할 수 있는 k의 선정 방법에 대하여도 추가적인 연구가 필요하며 추가 연구를 통하여 보다 정확한 k의 선정 방법을 제안할 수 있을 것으로 기대된다.

참고문헌

이재식, 박석두 (2007). 장르별 협업필터링을 이용한 영화 추천시스템의 성능 향상. <한국지능정보시스템학회논문지>, **13**, 65-78.

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

Ansari, A., Essegaier, S. and Kohli, R. (2000). Internet recommender systems. *Journal of Marketing Research*, **37**, 363-375.

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

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.

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

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

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.

Lekakos, G. and Giaglis, M. (2006). Improving the prediction accuracy of recommendation algorithms: Approaches anchored on human factors. *Interacting with Computers*, **18**, 410-431.

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. *Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work*, 175-186.

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'. *Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems*, 210-217.

Journal of the Korean Data & Information Science Society
2009, **20**(5), 809–818

한국데이터정보과학회지

A study on neighbor selection methods in k-NN collaborative filtering recommender system[†]

Seok Jun Lee[1]

[1]Department of MIS, Sangji University

Received 3 July 2009, revised 25 August 2009, accepted 12 September 2009

Abstract

Collaborative filtering approach predicts the preference of active user about specific items transacted on the e-commerce by using others' preference information. To improve the prediction accuracy through collaborative filtering approach, it must be needed to gain enough preference information of users' for predicting preference. But, a bit much information of users' preference might wrongly affect on prediction accuracy, and also too small information of users' preference might make bad effect on the prediction accuracy. This research suggests the method, which decides suitable numbers of neighbor users for applying collaborative filtering algorithm, improved by existing k nearest neighbors selection methods. The result of this research provides useful methods for improving the prediction accuracy and also refines exploratory data analysis approach for deciding appropriate numbers of nearest neighbors.

Keywords: Collaborative filtering, nearest neighbors selection, recommender system.

[†] This research was supported by a research fund in Sangji University (2008).
[1] Full Time Lecture, Department of MIS, Sangji University, Wonju 220-702, Korea.
E-mail: digitaldesign@sangji.ac.kr