협업 필터링 추천에서 대응평균 알고리즘의 예측 성능에 관한 연구

A study on the Prediction Performance of the Correspondence Mean Algorithm in Collaborative Filtering Recommendation

이 석 준 (Seok Jun Lee) 상지대학교 경영학과 겸임교수

이 희 춘 (Hee Choon Lee) 상지대학교 컴퓨터데이터정보학과 교수, 교신저자

요 약

본 연구의 목적은 좀 더 정확한 고객 선호도 예측을 위한 협업 필터링 알고리즘의 예측 성능을 평가하기 위한 것이다. 고객 선호도 예측의 정확도를 비교하기 위하여 이웃 기반의 협업 필터링 알고리즘과 대응평균 알고리즘에 의한 고객 선호도 예측의 MAE를 비교하였다. 예측 알고리즘의 정확성을 분석하기 위하여 MovieLens 1 Million dataset을 이용하여 실험을 하였다. 각 예측 알고리즘에 사용된 유사도 가중치는 일반적으로 이용되는 피어슨 상관계수와 벡터 유사도를 이용하였으며 분석결과 대응평균 알고리즘의 예측 정확도가 이웃 기반의 협업 필터링 알고리즘의 예측 정확도 보다 우수한 것으로 나타났다. 두 알고리즘에 사용된 유사도 가중치인 피어슨 상관계수와 벡터 유사도는 두 고객이 특정 상품에 대하여 공통으로 평가한 선호도 평가치를 이용하여 계산된다. 이때 공통으로 평가한 선호도 평가치의 개수가 적으면 계산된 유사도 가중치가 과대 평가된다. 과대 평가된 유사도 가중치를 보정하여 고객 선호도 예측의 정확도를 높이기 위하여 기존의 연구에서 고려한 공통 평가 영화의 개수 보다 확대된 범위를 적용하였으며 각 예측 방법에 따라 서로 다른 개선 경향을 파악할 수 있었다.

키워드 : 추천시스템, 협업 필터링, 유사도 가중치, 대응평균 알고리즘

Ⅰ. 서 론

최근 초고속 인터넷 인프라의 확산과 인터넷의 대중화로 다양한 형태의 전자상거래(e-commerce)가 활발하게 진행되고 있으며 인터넷 쇼핑뿐만 아니라 영화나 음악 서비스와 같은 다양한 형태의 인터넷 서비스가 급성장하고 있다. 특히 2006년 6월 현재 만6세 이상 인구 중 73.5%에 이르는 33,580천명이 '최근 1개월 이내 1회 이상' 인터넷을 이용한 것으로 나타났다. 인터넷 이용률은 2005년 6월에 비해 71.9%에서 1.6% 증가하였으며, 인터넷 이용자수로는 2005년 6월 32,570천명에서 1,010천명 증가하였다. 또한 만12세 이상 인터넷 이용자 중 최근 1년 이내에 인터넷을 이용하여 상품을 구매

하거나 예약/예매 등 인터넷쇼핑을 한 적이 있는 인터넷쇼핑 이용률은 51.3%로 2005년 6월 48.2%에 비해 3.1%p 증가하였다(2006년 상반기 정보화 실태조사 한국인터넷진흥원).

이러한 환경에서 전자상거래 기업들은 더욱 치열해진 경쟁에서 생존하기 위해 다양한 마케팅 전략을 구사하여야 하고 고객들도 기존의 서비스에 비해 차별화된 서비스를 원하고 있다. 전자상거래 기업들은 고객획득에서 고객유지의 마케팅 전략으로 고객에 대한 차별적인 서비스를 제공하기 위해 변하고 있다. 또한 전자상거래 기업들은 고객과의 상호관계를 향상시키기 위해 일대일 마케팅(one-to-one marketing),개인화(personalization), 고객화(customization)등의 서비스를 고객에게 제공함으로써 고객의 취향이나 관심에 초점을 맞춘 제품이나 서비스의 제공을 생존의 필수적인 전략으로 인식하고 있다.

본 연구는 추천시스템의 협업 필터링에서 고객간의 관계를 중심으로 선호도를 예측한 사용자 기반(user-based)의 예측 기법과 상품간의 관계를 중심으로 선호도를 예측한 아이템 기반(item-based)의 예측 기법으로 나누어 선호도 예측 알고리즘들의 예측 정확도를 비교 분석하고 선호도 예측의 정확도를 높이는 방법에 대하여 연구하였다. 분석을 위하여 본 연구에서는 GroupLens 연구소의 MovieLens dataset을 이용하여 영화에 대한 고객의 선호도를 예측하였으며, 영화에 대한 고객 선호도 예측을 위하여 GroupLens에서 제시한 최초의 자동화된 알고리즘인 이웃 기반의 협업 필터링(Neighborhood Based Collaborative Filtering) 알고리즘(Resnick 등, 1994)과 이를 개선한 대응 평균(Correspondence Mean)알고리즘(Lee, 2006)을 이용하여 고객 선호도를 예측하였다.

고객과 상품간의 선호도 유사 정도를 나타내는 유사도 가중치(similarity weight)는 피어슨 상관계수와 벡터 유사도를 이용하였으며, 예측의 정확도를 높이기 위하여 유사도 가중치 계산에 사용되는 공통 평가 쌍(co-rating)의 영향을 고려한 유의성 가중치(significance weight)를 기존의 연구에서 적용한 범위 보다 확대하여 선호도 예측 정확도 개선의 변화를 살펴보았다.

본 연구의 구성은 2장에서 문헌연구를 통하여 추천 시스템에 대한 기존의 연구를 살펴보고 3장 추천 알고리즘에서 본 연구의 실험에 사용된 선호도 예측 알고리즘인 이웃 기반의 협업 필터링 알고리즘(NBCFA)과 대응평균 알고리즘(CMA)에 대하여 살펴보고 각 알고리즘에 사용되는 유사도 가중치와 선호도 예측 정확도 평가 척도인 MAE, 그리고 공통 평가 쌍의 개수를 고려한 유사도 가중치의 보정치인 유의성 가중치에 대해 살펴본다. 4장에서는 실험 dataset의 구성과 연구방법에 대해 살펴보고 실험의 결과를 정리하였다. 5장에서는 본 연구의 결론을 유도하여 정리하였다.

Ⅱ. 이론적 배경

2.1 추천시스템의 개념

전자상거래는 고객에게 다양한 종류의 제품이나 서비스를 서로 비교하여 상대적으로 품질이 양호하고 저가인 제품을 시간과 장소에 제한을 받지 않고 구매할 수 있는 기회를 제공한다. 그러나 인터넷에서의 고객은 자신이 원하는 제품을 구매하기 위하여 많은 제품들 중에 자신의 취향에 맞는 제품이나 서비스를 스스로 찾아야 하는 정보 과부하(information overload) 현상에 직면하게 되고 구매 의욕까지 상실할 수도 있다. 이를 해결하기 위하여 고객에게 개인화된 서비스를 제공하는 시스템을 개인화 추천시스템(Personalized Recommender System)이라 부른다.

추천시스템은 전자상거래에서 고객의 입장과 기업의 입장을 모두 만족시킬 수 있는 시스템으로 받아들여지고 있다(Schafer, 2001). 추천시스템은 고객의 취향이나 선호도를 미리 예측하여 고객이 선호하리라 생각되는 상품을 미리 추천하여 줌으로써 고객에게는 정보탐색비용의 절감과 기

업에게는 판매상품에 대한 수요예측의 자료로 이용할 수 있으며 목표고객의 설정을 통한 마케팅 전략에도 활용할 수 있다(Riedl, 2002). 대표적으로 Amazon.com 등에서는 추천시스템을 이용하여 다양한 마케팅 전략을 구사하고 있다. 따라서 추천시스템에서 정확한 상품 추천을 위한 추천 알고리즘의 개발과 알고리즘을 통한 고객 선호도 예측의 활용이 중요해지고 있다. 또한 전자상거래의 규모가 확대됨에 따라 전자상거래를 이용하는 고객의 수와 온라인으로 판매되는 상품의 종류와 수가 늘어나고 있다. 이에 따라 고객과 상품의 규모 확장에 대응하는 대용량의 데이터베이스 환경에서 최대의 성능을 발휘할 수 있는 알고리즘의 필요성이 대두되고 있다.

2.2 추천시스템의 분류

추천시스템은 추천 방식에 따라 다양한 방식으로 나눌 수 있지만 일반적으로 내용기반(content-based)의 추천시스템과 협업 필터링(collaborative filtering)으로 나눌 수 있다. 내용기반의 추천시스템은 시스템 초기에 높은 정확도를 나타내는 것으로 알려져 있지만 상품에 대한 고객의 정보가 축적되면 협업 필터링의 정확도가 높아지며 두 가지 방식의 장점을 결합한 혼합(Hybrid)추천의 방식도 제안되었다(Claypool 등, 1999).

내용기반 추천시스템은 정보검색 분야의 기법을 이용하여 주로 상품의 특성에 대한 문자 내용을 분석하기 위해 적용된다. 상품의 추천은 고객이 이전에 구매하거나 경험한 상품의 특성에 대하여 고객의 선호 정도나 행동패턴과 같은 정보를 이용하며 추천 상품과 고객의 선호 정보 간의 일치 정도를 기반으로 상품의 추천이 이루어진다. 내용기반 추천기법은 초기의 추천시스템에서 우수한 성능을 발휘하였지만 다음과 같은 단점을 내포하고 있다(Balabanovic, Shoham, 1997; 김용수, 2006).

- 추천 상품의 속성은 반드시 문자로 이루어져야 하며 멀티미디어 파일을 직접적으로 추천할 수는 없다.
- 고객의 과거 경험만을 토대로 추천이 이루어지기 때문에 고객 본인의 선호도에만 특화된 상품만을 추천하고 뜻하지 않은 상품의 추천은 불가능하다.
- 상품의 품질이나 스타일, 유행, 혹은 상품에 대한 개인들의 견해 등을 반영한 추천을 할 수 없다.

협업 필터링 추천시스템은 내용기반의 추천시스템의 단점을 보완하여 현재 전자상거래에서 가장 성공적인 추천 기법으로 알려져 있으나 역시 몇 가지의 단점을 가지고 있다. 협업 필터링의 개념은 역사적으로 오래 전부터 적용되어 왔다. 협업 필터링은 단체나 집단의 구성원들이 어떤 것이 좋은 것이고 어떤 것이 나쁜 것인지를 협동적으로 알아내는 매카니즘이라고 할 수 있다. 선사 시대에는 이러한 매카니즘이 새로운 과일이나 곡식이 발견되면 모든 구성원들이 동시에 먹지 않고 누군가가 그것을 먹고 이상이 있는지를 확인한 후 안전하면 구성원들이 먹고 이것이 구전되어 전체 구성원들에게 알려졌던 것을 협업 필터링의 개념이라 할 수 있다. 인류의 예술, 철학, 과학 등이 이러한 경험적인 과정을 통하여 축적되었다고 할 수 있다. 협업 필터링과 동일한 과정이 우리의 시간과 부에 가치가 있는 이론과 창작이 무엇인지를 결정하는데 도움을 주고 있다(Riedl, Konstan, 2002).

2.3 협업 필터링(Collaborative filtering)

협업 필터링 추천기법은 고객의 개인적 취향, 개인 정보 등과 같은 부가적인 정보와 상품이 가지고 있는 특성과 같은 복잡한 정보를 의도적으로 무시하고 고객-상품 간의 상호관계만을 이용하는 접근법을 취하는데 일반적으로 이러한 관계는 상품에 대한 고객의 선호도 값과 같이 간단한 형태의 관계 데이터로 구성되어 있다(Hill 등,

1995; Resnick 등, 1994; Shardanand, Maes, 1995). 협업 필터링에서 고객과 상품 간의 상호관계 데이터는 명시적 데이터와 암시적 데이터로 구분되며 암시적 데이터는 웹사이트에 접속한 고객의 로그 분석을 통해 상품에 대한 선호도를 유추하며 반면 명시적 자료는 고객이 직접 평가한 선호도 평가치를 이용한다. 협업 필터링 추천의 간단한 예로 모든 고객들에게 가장 잘 알려진 상품을 추천하는 것을 들 수 있다. 그러나 이 방법은 고객 본인의 선호도를 반영하지 못한다. 일반적으로 협업 필터링에서 고객과 상품 간의 상호관계 데이터는 행렬 형태로 표현하고 분석한다. 협업 필터링은 가장 널리 이용되고 성공적인 추천 접근법으로 알려져 있으며 추천 알고리즘 연구의 근간을 이루고 있다(Breese 등, 1998; Resnick 등, 1994; Resnick 등, 1997).

협업 필터링에서 다양한 범주의 알고리즘들이 제안되었다. 대부분의 협업 필터링 추천시스템은 사용자 기반(user-based)의 협업 필터링(Burke, 2000; Claypool 등, 1999; Mobasher, 등, 2000; Sarwar 등, 1998)과 아이템 기반(item-based)의 협업 필터링(Deshpande, Karypis, 2004; Sarwar 등, 2001)으로 나누어진다. 사용자 기반의 협업 필터링은 추천을 하고자 하는 고객과 이웃 고객들 간에서 얻어진 상호관계 데이터를 이용하고 아이템 기반의 협업 추천기법은 상품들 간의 상호 유사관계를 이용하여 추천이 이루어진다. 현재 전자상거래 추천시스템은 일반적으로 아이템 기반의 협업 필터링 알고리즘을 이용한다(Schafer 등, 2001). 유사 이웃의 구성, 분류를 위한 알고리즘, 연관규칙 마이닝, 베이지안 네트워크, 그리고 군집화 모형과 같은 많은 데이터 분석 알고리즘들이 협업 필터링 문제에 적용되었다. 김경재, 김병국 (2005)은 유전자 알고리즘을 이용하여 고객 정보에서 성향을 추출할 수 있는 추천시스템의 추천엔진 개발에 대해 연구하였다. 거래기록에서 상품과 고객 간의 연관규칙에 대한 패턴을 끌어내기 위하여 연관규칙 마이닝 기법이 추천시스템에 적용되었다(Lin 등, 2002; 김재경 등, 2003). 베이지안 네트워크는 상품에 대한 고객의 경험들 간에 종속관계를 수학적으로 표현하는데 이 종속관계에는 상품에 대한 선호도의 인과관계와 상관관계가 반영되어 있다(Heckerman 등, 2001). 또한 K-means 군집화 알고리즘을 이용하여 거리개념의 군집을 생성한 후 군집간의 순차적 패턴을 발견하기 위한 방법도 연구되고 있으며 이와 같은 생성적 군집화 모형을 근거로 추천이 이루어진다(심장섭, 2005).

Ⅲ. 추천 알고리즘

3.1 이웃 기반의 협업 필터링 알고리즘(Neighborhood-Based Collaborative Filtering Algorithm)

전자상거래에 적용되고 있는 협업 필터링의 개념은 Xerox Palo Alto 연구소에서 메일의 분류를 위한 실험 시스템인 Tapestry에 의해 소개되었다(Goldberg 등, 1992). GroupLens에서는 인터넷을 기반으로 형성된 토론 시스템인 유즈넷 뉴스(UseNet News) 그룹의 기사를 추천하기 위해 최초의 자동화된 협업 필터링 알고리즘인 이웃 기반의 협업 필터링(Neighborhood-Based Collaborative Filtering) 알고리즘을 제안하였다(Resnick 등, 1994). 초기 GroupLens 시스템에서는 고객들 간의 선호도 유사 정도를 나타내기 위한 유사도 가중치(similarity weight)로 피어슨 상관계수(pearson's correlation coefficient)를 이용하였고 유즈넷 뉴스 그룹의 모든 고객들의 선호 기사에 대한 상관관계를 이용하였다. 특정 문서에 대하여 추천을 받을 고객의 선호도를 예측하기 위해 다음의 NBCFA를 이용하여 최종 선호도 예측을 계산하였다.

$$\widehat{U_x} = \overline{U} + \frac{\sum_{J \in Raters} (J_x - \overline{J}) r_{uj}}{\sum_{J \in Raters} |r_{uj}|} \quad (1)$$

식(1)에서 $\widehat{U_x}$는 상품 x에 대한 특정 고객 u

가 어느 정도의 선호도를 보일 것인지에 대한 선호도 예측치이고, $\overline{U}$는 특정 고객 u가 선호도를 평가한 모든 상품에 대한 선호도 평균이며, J_x 는 고객 j가 평가한 실제 선호도 평가치이고, r_{uj}는 특정 고객 u와 이웃한 고객 j의 유사 정도를 나타내는 유사도 가중치이고, $\overline{J}$는 이웃 고객 j가 평가한 모든 상품에 대한 선호도 평균이다. *Raters*는 상품에 대해 선호도를 표시한 고객들을 의미한다. 여기서 특정 고객은 상품을 추천 받거나 혹은 추천을 하기 위한 고객을 의미한다.

3.2 유사도 가중치(similarity weight)

NBCFA을 적용하기 위한 첫 번째 단계는 특정 상품에 대한 선호도 예측을 위해 이웃 고객을 정하고 이웃 고객과 추천을 받을 고객 간의 선호도 유사 정도를 나타내는 유사도 가중치를 구하는 것이다. 특정 상품을 추천 받고자 하는 고객들은 이미 그룹 내의 이웃 고객들이 경험하여 얻어진 정보를 바탕으로 보다 정확한 추천을 받기를 원한다. 이때 특정 상품을 평가한 이웃 고객과 추천을 받고자 하는 고객과의 선호도 유사 정도를 계량적으로 나타내는데 이때의 선호도 유사 정도를 나타내는 값을 유사도 가중치(similarity weight)라 하며 최초의 GroupLens 시스템에서는 피어슨 상관계수가 이용되었다 (Resnick 등, 1994).

다음은 고객 와 이웃 고객 의 선호도 유사 정도를 나타내는 유사도 가중치인 피어슨 상관계수이다.

$$r_{uj} = \frac{\sum_{i=1}^{m}(R_{u,i} - \overline{R_u})(R_{j,i} - \overline{R_j})}{\sqrt{\sum_{i=1}^{m}(R_{u,i} - \overline{R_u})^2 \cdot \sum_{i=1}^{m}(R_{j,i} - \overline{R_j})^2}} \quad (2)$$

r_{uj}는 추천을 받고자 하는 고객 u와 그룹 내의 이웃 고객 j가 선호도를 평가한 상품에 대한 피어슨 상관계수이며 이를 두 고객 간의 유사도 가중치로 이용한다. 여기서 $R_{u,i}$는 고객 u가 선호도를 평가한 상품 i의 선호도 평가치이고, $\overline{R_u}$는 고객 u가 선호도를 평가한 상품들의 선호도 평가치의 평균이며, $R_{j,i}$는 이웃 고객 j가 선호도를 평가한 상품 i의 선호도 평가치(rating)이고, $\overline{R_j}$는 이웃 고객 j가 선호도를 평가한 상품들의 선호도 평가치의 평균이다.

또한 두 문서간 유사성을 계산을 위해 각 문서에서 단어의 출현 빈도를 벡터로 처리하여 계산하는데 이때의 코사인 벡터를 협업 필터링에 적용하여 이웃 고객과의 선호도 유사 정도인 유사도 가중치로 사용하며 이를 벡터 유사도(vector similarity)라 한다(Breese 등, 1998). 피어슨 상관계수와 벡터 유사도는 모두 고객 간의 선호도 유사 정도를 나타내며 피어슨 상관계수는 두 고객의 선호도 유사 정도를 1에서 -1까지의 양과 음의 관계로 표현하는 반면 벡터 유사도의 경우는 음의 값이 존재하지 않으며 최대 1의 유사도 가중치 값으로 정의된다. 다음은 고객 u와 이웃 고객 j의 선호도 유사 정도를 나타내는 유사도 가중치인 벡터 유사도이다.

$$r_{uj} = \cos(\overrightarrow{R_u} \cdot \overrightarrow{R_j}) = \frac{R_u \cdot R_j}{|R_u| \cdot |R_j|} \quad (3)$$

Breese 등(1998)은 피어슨 상관계수와 벡터 유사도 외에 사용할 수 있는 유사도 가중치로 기본 선호도(default voting), 역사용자 빈도(inverse user frequency), 사례확대(case amplification) 등의 다양한 유사도 가중치를 소개하고 평가하였다.

본 연구에서는 피어슨 상관계수와 벡터 유사도를 고객 간의 선호 정도를 나타내는 유사도 가중치로 이용하여 분석하였다.

3.3 대응평균 알고리즘(Correspondence Mean Algorithm)

GroupLens에서 제시한 NBCFA를 수정한 대

응평균 알고리즘(Correspondence Mean Algorithm)은 100K MovieLens dataset을 분석한 결과 NBCFA의 결과보다 향상된 예측력을 보였다(Lee, 2006). Herlocker 등 (2004)의 연구에서 5점 척도의 선호도 자료에서 NBCFA의 MAE가 0.73의 자연적 장애물인 *Magic Barrier*에 도달한다고 밝히고 있으며 선호도 예측력을 높이기 위한 다양한 방법을 이용하여도 0.73이하로 낮추기 어려움을 지적하고 있다(Herlocker 등, 2004). 그러나 CMA의 경우 0.73 이하의 MAE를 나타내고 있다(Lee, 2006). NBCFA을 수정한 CMA는 수식(4)와 같다.

$$\widehat{U}_x = \overline{U}_{match} + \frac{\sum_{J \in Raters} (J_x - \overline{J}_{match}) r_{uj}}{\sum_{J \in Raters} |r_{uj}|} \quad (4)$$

GroupLens에서 제시한 NBCFA에서 $\overline{U}$는 추천을 받고자 하는 고객 u의 선호도 평가치 전체의 평균을 사용하고 있다. 이때의 $\overline{U}$는 고객 u 자신의 선호도를 나타낸다. 그러나 특정 고객 u와 이웃 고객 j의 유사 정도를 나타내는 유사도 가중치 r_{uj}는 고객 u와 이웃 고객 j가 공통으로 평가한 상품의 선호도 평가치들로 계산된다. 여기서 $\overline{U}$를 고객 u가 평가한 영화 전체의 선호도 평가치의 평균을 이용하면 고객 u 자신의 선호도가 과대평가되어 고객 j의 선호도를 충분히 반영하지 못하게 된다. 그래서 $\overline{U}$를 고객u가 평가한 영화 전체의 평균이 아닌 고객 u와 이웃 고객 j가 공통으로 평가한 상품의 선호도 평가치들의 평균들을 계산하고 다시 이 평균들의 평균을 구한 $\overline{U}_{match}$ 로 정의하였다. 또한 고객 u와 이웃 고객 j가 공통으로 평가한 상품들의 선호도 평가치를 이용하여 고객 u의 선호도를 예측하기 때문에 이웃 고객 j의 선호도인 $\overline{J}$도 고객 j가 평가한 영화 전체의 평균을 이용하면 고객 j 자신의 선호도가 과대평가된다. 그래서 고객 j의 선호도를 나타내는 $\overline{J}$도 동일한 방법으로 $\overline{J}_{match}$로 정의하였다.

3.4 선호도 예측 정확도 평가척도

MAE(Mean Absolute Error)는 협업 필터링에 의한 예측치의 성능을 평가하기 위해 가장 일반적으로 적용되는 평가 척도이다. MAE는 계산된 선호도 예측치와 이에 대응하는 실제 선호도 평가치의 절대 편차의 평균으로 계산된다(Breese 등, 1998; Shardanand, Maes, 1995).

$$MAE = \frac{1}{N}\sum_{j=1}^{N} |R_{uj} - \hat{R}_{uj}| \quad (5)$$

여기서, N은 추천을 받을 모든 고객들에 대한 예측의 총 개수를 나타내며 R_{uj}는 실제 $\hat{R}_{uj}$ 선호도 평가치이고 는 R_{uj}에 대응하는 예측치이다. MAE에 의한 성능평가의 결과는 MAE가 낮을수록 전체 예측 알고리즘의 정확도가 높다. MAE와 유사한 평가 척도로 MSE(Mean Squared Error), RMSE(Root Mean Squared Error), 그리고 MAE를 표준화 시킨 NMAE(Normalized Mean Absolute Error)등이 있으며 일반적으로 전체 시스템의 정확도는 MAE를 이용하여 성능을 평가한다.

3.5 공통 선호도 평가 쌍(co-ratings)의 영향

Herlocker 등(2002)은 고객간의 선호도 유사관계를 나타내는 유사도 가중치인 상관계수가 두 고객이 공통으로 선호도를 표기한 공통 평가 쌍의 개수에 영향을 받고 있음을 연구하였으며 공통 평가 쌍의 개수가 50개 이하일 경우 두 고객의 선호도 유사 정도가 과대 평가되어 선호도 예측의 오차가 커지며 과대 평가된 유사도 가중치를 줄여 줌으로써 선호도 예측이 개선됨을 보였다. 또한 유사도 가중치에 영향을 미치는 공통 응답 쌍의 개수를 50개로 보았으며 50개 이하의 공통 응답 쌍을 갖는 유사도 가중치는 선호도 예측 정확도에 영향을 미치며 그 이상의 쌍의 개수를 갖는 경우는 선호도 예측의 정확도에 영향을 미치

지 않기 때문에 50개로 설정하였다. 공통 평가 쌍이 50개 이하인 유사도 가중치를 줄여주기 위한 보정치가 유의성 가중치(significance weight)이며 공통 평가 쌍의 개수가 50개 이하일 경우 유사도 가중치에 n/50의 보정치인 유의성 가중치를 곱하여 유사도 가중치의 영향을 줄여주었다. 여기서 n은 두 고객의 공통 평가 쌍의 개수이며 n이 50보다 클 경우는 유의성 가중치를 1로 설정하여 유사도 가중치 본래의 값을 그대로 사용하였다. 공통 평가 쌍의 영향에 대하여는 Lee 등(2006)에서 100K MovieLens dataset에 대하여 연구하였고 손재봉, 서용무(2006)의 연구에서도 DOM을 정의하여 Million dataset에 대하여 연구하였다. 손재봉, 서용무 (2006)의 연구에서는 50개까지 적용하여 연구하였으며 이희춘, 이석준(2006)의 연구에서는 공통 평가 쌍의 개수를 300개까지 확대하여 적용하였다.

벡터 유사도의 경우도 두 고객이 공통으로 평가한 영화의 선호도 평가치로 계산되기 때문에 본 연구에서는 유의성 가중치를 동일하게 적용시켜 보았다. 그러나 본 연구에서는 기존의 연구와 달리 사용자 기반의 예측뿐만 아니라 아이템 기반의 예측도 실험하였다.

$$U_x = \overline{U} + \frac{\sum_{J \in Raters} (J_x - \overline{J}) r'_{uj}}{\sum_{J \in Raters} |r'_{uj}|}$$

여기서, $r'_{uj} = r_{uj} \cdot sw$ (6)

sw는 과대 평가된 유사도 가중치의 영향을 줄여주기 위한 유의성 가중치로 다음과 같이 설정한다.

$$sw = \frac{\min(n(I_u \cap I_j), C)}{C} \quad (7)$$

여기서 $n(I_u \cap I_j)$는 고객 u와 j가 공통으로 평가한 영화 편수이고, C는 유사도 가중치에 영향을 미치는 공통 평가 쌍의 개수의 결정하기 위한 한계치로 기존의 연구에서는 일반적으로 50으로 설정하였다.

Ⅳ. 실험설계 및 분석

4.1 실험 dataset의 구성

MovieLens dataset은 GroupLens 연구 프로젝트에 의해 수집된 영화에 대한 평가 자료로 100K, 1 million, 미공개 진행 dataset으로 나누어져 있다. 100K dataset은 943명의 인원이 1682편의 영화에 대해 평가한 자료로 10,000개의 선호도 평가치로 구성되어 있으며 Million dataset은 6040명의 평가 인원이 3952편의 영화에 대한 평가로 구성되어 있고 실제 dataset에서 246편의 영화는 평가되지 않았으며 평가치의 개수는 1,000,209개로 구성되어 있다. 미공개 dataset은 현재 자료가 수집 중이며 21,526명의 평가 인원이 8,848편의 영화에 대한 평가로 구성되어 있으며 2,933,690개의 평가치로 구성되어 있다. 또한 간단한 인구 통계 정보와 함께 평가 영화에 대한 정보를 제공하고 있다. 본 연구는 MovieLens 1 million dataset을 80%의 훈련집합(training set)과 20%의 실험집합(test set)으로 분할한 3개의 실험 dataset으로 알고리즘의 예측력을 평가하였다.

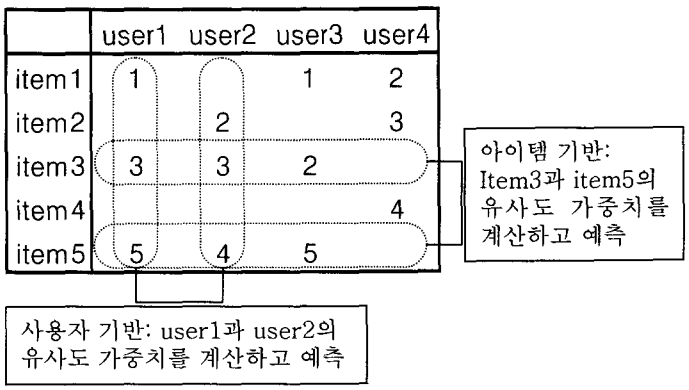
4.2 실험 방법

다음 <그림2>는 본 연구의 실험 방법으로 80%의 훈련집합과 20%의 실험집합으로 구성된 3개의 실험 dataset에 대하여 사용자 기반의 예측과 아이템 기반의 예측으로 나누어 진행된 실험의 흐름도이다. 본 연구는 식(1), 식(4)에서 소개된 NBCFA, CMA를 이용하여 고객 선호도를 예측하였으며, 각 알고리즘에 고객과 상품의 유사관계를 나타내기 위한 유사도 가중치를 피어슨 상관

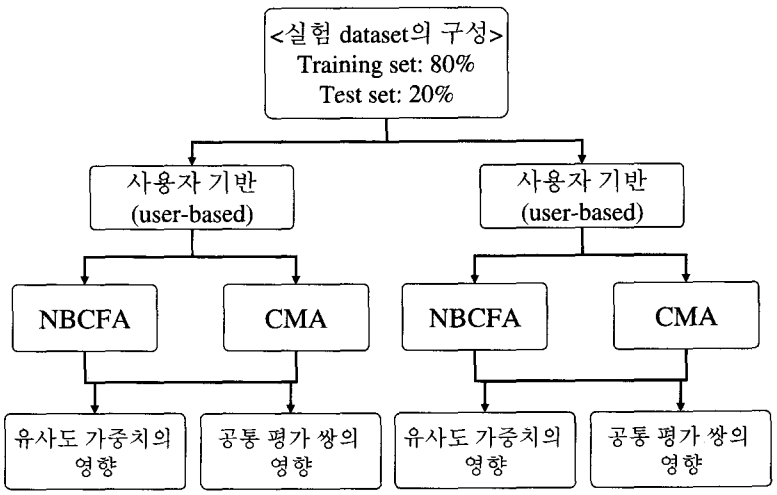
계수와 벡터 유사도로 나누어 적용시켜 실험하였다. 유사도 가중치인 피어슨 상관계수와 벡터 유사도는 사용자 기반의 예측의 경우 고객간의 선호도 유사 정도를 나타내며 두 사용자가 각각 표기한 선호도 평가치 중 동일 상품에 대하여 공통으로 평가한 선호도 평가치(co-rating)를 이용하여 계산된다. 이 때 전술하였듯이 공통으로 평가한 쌍(co-rating)의 개수가 고객간의 선호도 유사 관계에 영향을 미치는 것으로 알려져 있으며 기존의 연구에서는 공통으로 평가한 영화의 편수가 50편 이하일 경우 고객간의 유사 정도가 과대하게 평가되는 것으로 분석되었고 이 영향을 줄여주기 위한 방법으로 유의성 가중치를 설정하여 유사도 가중치의 영향을 줄여 주었다(Herlocker 등, 2002; 손재봉, 서용무, 2006; 이희춘, 이석준, 2006). 아이템 기반의 예측에서는 이 유사도 가중치가 고객간의 관계를 나타내는 것이 아니라 아이템, 즉, 상품들의 유사 관계를 나타내는 가중치이다.

<그림 1>은 사용자 기반의 예측과 아이템 기반의 예측을 도식화한 것이다.

사용자 기반의 예측에서 설정한 유의성 가중치가 아이템 기반의 예측에도 영향을 줄 것



〈그림 1〉 사용자 기반과 아이템 기반의 예측



〈그림 2〉 연구흐름도

으로 판단하여 동일한 방법으로 유사도 가중치에 공통 평가 쌍의 영향을 고려하였으며 사용자 기반의 예측과 아이템 기반의 예측에서 공통 평가 쌍의 개수의 분포를 분석하여 기존의 연구에서 설정하였던 범위보다 확장된 범위를 실험에 적용시켰다.

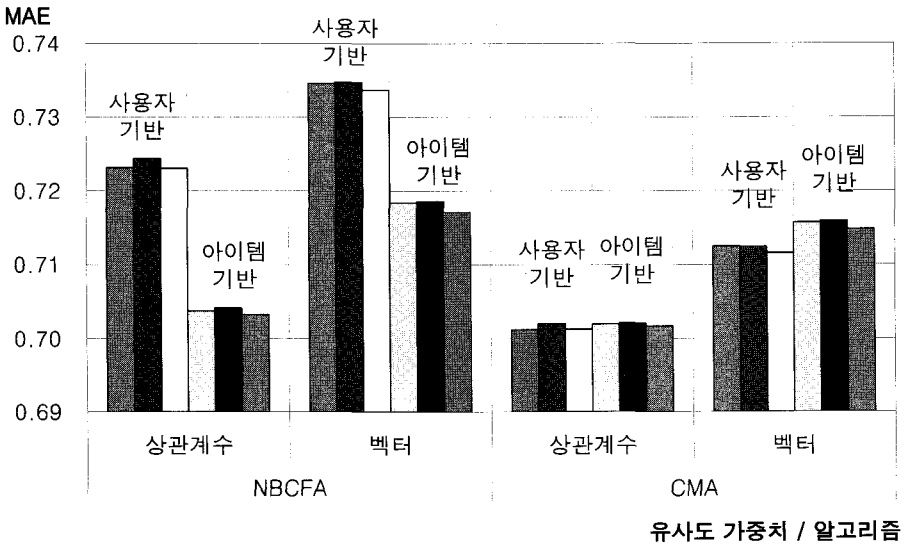
4.3 분석 결과

4.3.1 유사도 가중치에 따른 알고리즘 별 예측 정확도

다음 <표 1>과 <그림 3>은 사용자 기반의 예측과 아이템 기반의 예측 방식에 따라 유사도 가중치에 따른 각 알고리즘의 결과표이다. 결과에서 먼저 유사도 가중치에 의한 결과는 각 예측 알고리즘에 대하여 사용자 기반의 예측과 아이템 기반의 예측에서 모두 피어슨 상관계수의 유사도 가중치를 이용한 경우의 선호도 예측 정확도가 벡터 유사도를 유사도 가중치를 이용한 경우보다 우수함을 알 수 있다. 또한 사용자 기반의 예측 결과와 아이템 기반의 예측 결과를 비교하면 NBCFA에서는 아이템 기반의 예측 결과가 우수하게 나타났으나 CMA의 경우 사용자 기반의 예측 결과가 우수한 것으로 분석되었다. 모든 실

〈표 1〉 사용자 기반과 아이템 기반의 예측에서 알고리즘과 유사도 가중치에 따른 선호도 예측 결과 MAE

알고리즘	유사도 가중치	사용자 기반			아이템 기반		
		dataset1	dataset2	dataset3	dataset1	dataset2	dataset3
NBCFA	상관계수	0.7231	0.7243	0.7230	0.7037	0.7042	0.7033
	벡터	0.7345	0.7347	0.7337	0.7184	0.7185	0.7171
CMA	상관계수	0.7012	0.7020	0.7013	0.7020	0.7021	0.7017
	벡터	0.7125	0.7124	0.7116	0.7157	0.7159	0.7148



〈그림 3〉 사용자 기반과 아이템 기반의 예측에서 알고리즘과 유사도 가중치에 따른 결과 그림

〈표 2〉 사용자 기반과 아이템 기반의 예측에서 예측 알고리즘과 유사도 가중치에 따른 개인별 선호도 예측 결과를 이용한 프리드만 검정(Friedman Test) 결과

실험 dataset	방식	유사도 가중치	알고리즘	평균	표준편차	평균순위	카이제곱	유의확률
dataset1	사용자 기반	상관계수	NBCFA	0.7465	0.2395	2.67	1223.39	0.000
			CMA	0.7285	0.2342	2.09		
		벡터	NBCFA	0.7483	0.2390	2.86		
			CMA	0.7317	0.2322	2.38		
	아이템 기반	상관계수	NBCFA	0.7171	0.2274	2.17	2038.20	0.000
			CMA	0.7164	0.2272	2.09		
		벡터	NBCFA	0.7345	0.2342	2.93		
			CMA	0.7328	0.2348	2.81		
dataset2	사용자 기반	상관계수	NBCFA	0.7510	0.2400	2.67	1199.43	0.000
			CMA	0.7323	0.2328	2.09		
		벡터	NBCFA	0.7524	0.2395	2.85		
			CMA	0.7351	0.2315	2.39		
	아이템 기반	상관계수	NBCFA	0.7212	0.2269	2.18	1905.36	0.000
			CMA	0.7207	0.2273	2.10		
		벡터	NBCFA	0.7384	0.2341	2.92		
			CMA	0.7371	0.2353	2.80		
dataset3	사용자 기반	상관계수	NBCFA	0.7479	0.2391	2.69	1138.59	0.000
			CMA	0.7301	0.2315	2.12		
		벡터	NBCFA	0.7501	0.2405	2.84		
			CMA	0.7315	0.2313	2.36		
	아이템 기반	상관계수	NBCFA	0.7180	0.2258	2.22	1638.48	0.000
			CMA	0.7177	0.2249	2.18		
		벡터	NBCFA	0.7501	0.2405	3.01		
			CMA	0.7315	0.2313	2.59		

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

험 dataset에서 CMA의 예측 정확도가 NBCFA의 결과보다 우수한 것을 알 수 있다.

<표 1>에서 제시된 MAE는 실험 dataset 전체의 선호도 예측 오차를 이용한 MAE이며 시스템 전체의 정확도를 평가하기 위하여 일반적으로 사용된다. 그러나 본 논문에서는 알고리즘의 예측 정확도를 통계적으로 분석하기 실제 선호도 평가치와의 예측 오차를 개인별로 정리하고 각 개인 별 선호도 예측 오차의 평균인 개인별 MAE를 구하였다. <표 2>는 각 알고리즘의 개인별 MAE 결과를 이용한 프리드만 검정(Friedman Test)의 결과이다. 모든 실험 dataset에서 유

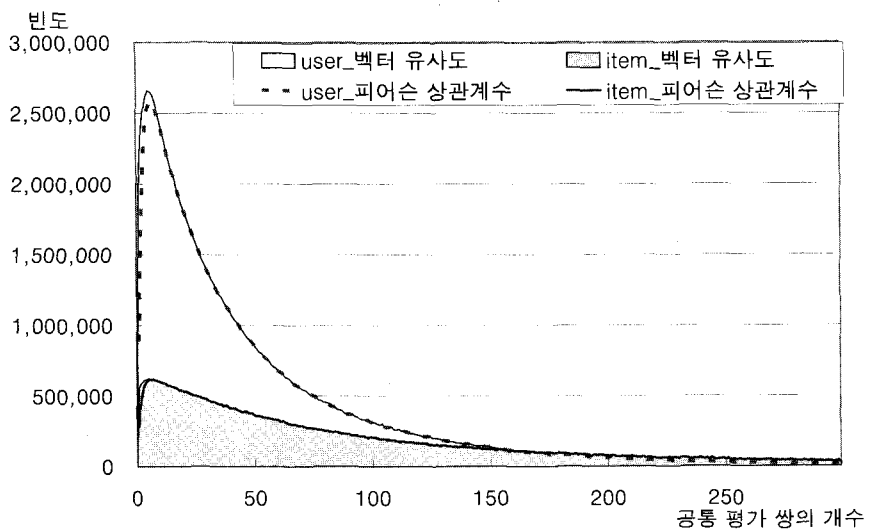
의수준 $\alpha = 0.001$에서 통계적으로 유의한 결과를 얻었으며 결과에서 CMA의 결과가 NBCF에 비하여 성능이 우수함을 알 수 있고 피어슨 상관계수를 유사도 가중치로 사용한 CMA의 예측 결과가 가장 우수한 예측력을 가짐을 알 수 있다. 실험 dataset1의 결과를 보면 사용자 기반의 예측에서 상관계수를 유사도 가중치로 이용한 NBCFA의 개인별 MAE의 평균순위는 2.67이며 CMA의 평균순위는 2.09를 나타내며 벡터 유사도를 유사도 가중치로 이용한 NBCFA의 평균순위는 2.83, CMA는 2.38로 나타나 상관계수를 유사도 가중치로 이용한 CMA의 예측 정확도가 가장 높음을 통계적을 검증하였다. 이후 모든 실험 dataset과 예측 방식에 따라서도 동일한 경향의 결과를 얻고 있음을 알 수 있다.

4.3.2 공통 평가 응답 쌍(co-ratings)의 영향 설정

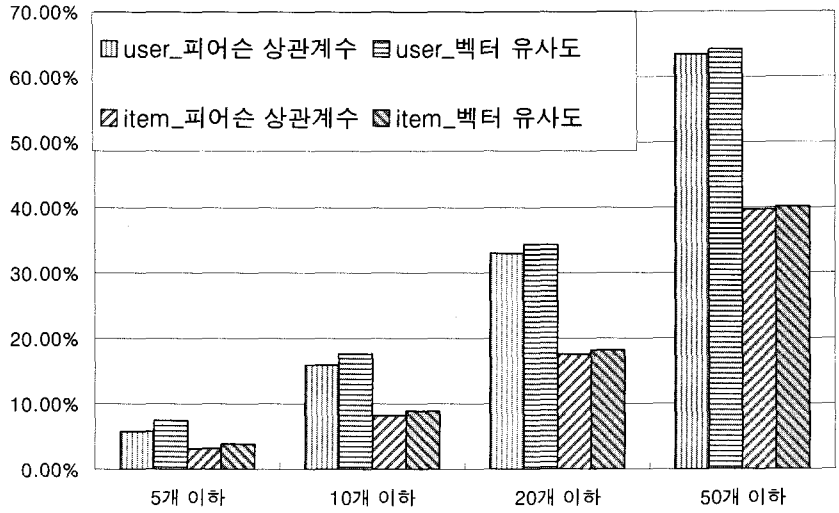
사용자 기반에서 계산되는 유사도 가중치와 아이템 기반에서 계산되는 유사도 가중치는 서로 다르게 계산되기 때문에 공통 평가 쌍의 분포가 서로 다르게 나타난다. 다음 <그림 4>는 본 연구에 사용된 실험 dataset1에서 사용자 기반의 예측에서 유사도 가중치를 구하기 위한 공통 평가 쌍의 분포와 아이템 기의 예측에서의 공통 평가 쌍의 분포이다.

<그림 4>와 <그림 5>에서 특정 고객과 이웃 고객이 공통으로 평가한 평가 쌍의 분포를 보면 평가 쌍의 개수가 10개 이하일 경우에 가장 높은 빈도를 나타남을 알 수 있으며 이는 공통으로 평가한 평가 쌍의 개수가 작아 유사도 가중치를 과대 평가할 가능성이 높음을 암시한다. 사용자 기반과 아이템 기반에서 동일한 경향의 분포가 나타남을 알 수 있다. 또한 공통 평가 쌍의 개수의 비율을 보면 50개 이하일 경우가 사용자 기반의 경우 전체에서 60% 가량을 차지함을 알 수 있지만 아이템 기반의 경우는 전체에서 40% 정도의 비율을 차지함을 알 수 있다. 본 연구는 이와 같은 분석결과를 토대로 유의성 가중치의 범위를 기존의 연구보다 확대하여 범위를 다음과 같이 설정하였다.

$$C = \{3,5,7,10,15,\cdots,50,60,\cdots,100,120,150,180,200,300,400,500,700,1000,2000,4000,7000,10000\}$$



〈그림 4〉 실험 dataset1에서 사용자 기반과 아이템 기반의 공통 평가 쌍의 개수 분포



〈그림 5〉 dataset1에서 공통 평가 쌍의 개수에 따른 누적비율

본 연구를 통하여 기존의 연구와 마찬가지로 공통 평가 응답 쌍이 선호도 예측 정확도에 영향을 미침을 알 수 있었으며 피어슨 상관계수뿐만 아니라 벡터 유사도에서도 공통 평가 쌍의 영향을 관찰 할 수 있었다. 실험 dataset에 따라 최소의 MAE를 갖는 유의성 가중치는 다르게 나타났지만 유의성 가중치에 따라 선호도 예측 정확도의 개선 경향은 유사하게 나타났다. 다음 <표 3>과 <그림 6>은 사용자 기반의 선호도 예측 결과로 각 실험 dataset별 유의성 가중치에 따라 피어슨 상관계수와 벡터 유사도의 최대 MAE와 최소 MAE에 대한 요약이다. 실험 dataset1의 결과를 보면 CMA의 MAE가 가장 우수하게 나타났으며 최소의 MAE를 갖게 하는 유의성 가중치는 n/120으로 분석되었다. Dataset1의 결과를 보면 피어슨 상관계수를 이용한 NBCF는 n/500에서 최소값인 0.7133을 보이며 CMA의 경우 n/120에서 0.6964의 최소 MAE를 보임을 알 수 있다. 또한 벡터 유사도의 겨우 n/700과 n/500에서 각각 0.7272과 0.7088의 MAE를 보임을 알 수 있으며 실험 dataset2와 3의 결과도 <표 3>에 정리하였다.

다음 <표 4>와 < 그림 7>은 아이템 기반의 예측에서 유의성 가중치에 따른 결과이다.

<표 4>와 <그림 7>에서 아이템 기반에서 응답 쌍의 개수의 영향은 사용자 기반의 결과와 다른 감소 형태를 보이고 있으며 예측 정확도가 안정화되는 유의성 가중치의 설정치도 훨씬 커짐을 알 수 있다. 사용자 기반의 예측결과와 비교하여 유의성 가중치에 따른 감소량도 사용자 기반의 경우보다 작음을 알 수 있다.

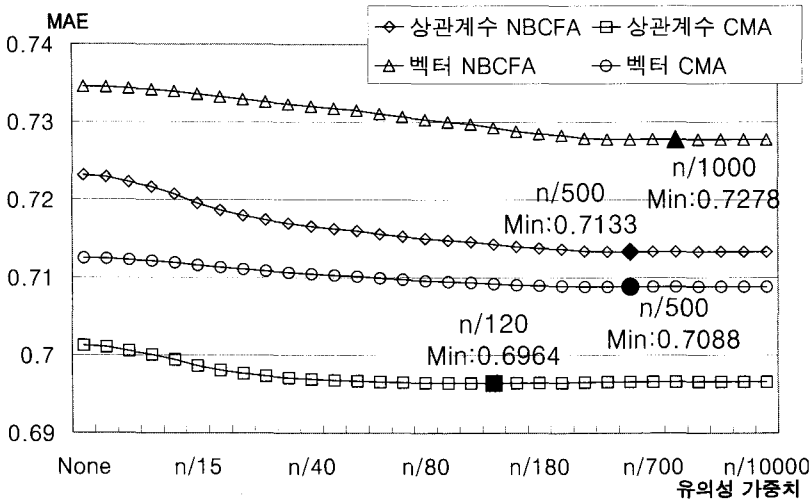
V. 결 론

본 논문은 협업 필터링에서 사용자 기반의 예측과 아이템 기반의 예측으로 나누어 MovieLens 1 million dataset을 분석하였다. 사용자 기반의 예측에서 특정 고객과 이웃 고객간의 선호도 유사관계와 아이템 기반의 예측에서 특정 상품과 이웃 상품간의 유사 관계를 나타내는 유사도 가중치로 피어슨 상관계수와 벡터 유사도로 나누어 NBCFA와 CMA에 각각 적용하여 선호도 예측의 정확도를 비교하였다.

분석결과 먼저 각 알고리즘에 대하여 유사도 가중치의 영향은 벡터 유사도의 결과보다 피어

〈표 3〉 사용자 기반의 예측에서 유의성 가중치에 따른 최대/최소 MAE 결과

실험 dataset	피어슨 상관계수			벡터 유사도		
	알고리즘	Max/Min	유의성 가중치	알고리즘	Max/Min	유의성 가중치
dataset1	NBCFA	0.7231	미적용	NBCFA	0.7345	미적용
		0.7133	n/500		0.7278	n/700
	CMA	0.7012	미적용	CMA	0.7125	미적용
		0.6964	n/120		0.7088	n/500
dataset2	NBCFA	0.7243	미적용	NBCFA	0.7347	미적용
		0.7143	n/4000		0.7278	n/10000
	CMA	0.7020	미적용	CMA	0.7124	미적용
		0.6970	n/100		0.7085	n/1000
dataset3	NBCFA	0.7230	미적용	NBCFA	0.7337	미적용
		0.7132	n/700		0.7270	n/700
	CMA	0.7013	미적용	CMA	0.7116	미적용
		0.6964	n/100		0.7079	n/500

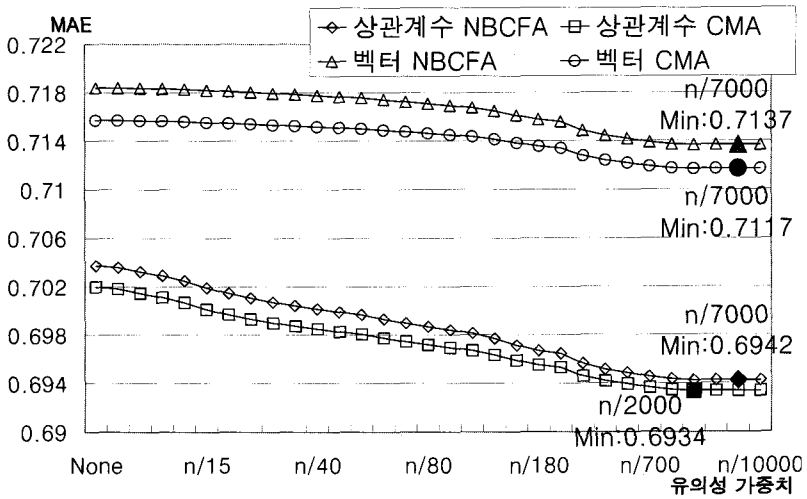


〈그림 6〉 사용자 기반의 예측에서 유의성 가중치에 따른 알고리즘과 유사도 가중치 별 MAE의 변화

슨 상관계수를 이용한 선호도 예측 결과가 우수함을 알 수 있다. 선호도 예측 알고리즘은 CMA의 예측 성능이 우수함을 알 수 있는데 이는 NBCFA에서 선호도 예측을 하는 특정 고객 자신의 선호도를 본인이 상품에 대해 평가한 전체 선호도의 평균을 이용할 경우 이웃 고객과의 관계가 고려되지 않아 선호도 예측의 오차가 커질 수 있음을 보여준다. 결국 NBCFA는 이

〈표 4〉 아이템 기반의 예측에서 유의성 가중치에 따른 최대/최소 MAE 결과

실험 dataset	피어슨 상관계수			벡터 유사도		
	알고리즘	Max/Min	유의성 가중치	알고리즘	Max/Min	유의성 가중치
dataset1	NBCFA	0.7037	미적용	NBCFA	0.7184	미적용
		0.6942	n/7000		0.7137	n/7000
	CMA Type2	0.702	미적용	CMA Type2	0.7157	미적용
		0.6934	n/2000		0.7117	n/7000
dataset2	NBCFA	0.7042	미적용	NBCFA	0.7185	미적용
		0.6946	n/4000		0.7135	n/10000
	CMA Type2	0.7021	미적용	CMA Type2	0.7159	미적용
		0.6934	n/7000		0.7115	n/7000
dataset3	NBCFA	0.7033	미적용	NBCFA	0.7171	미적용
		0.6938	n/7000		0.7122	n/4000
	CMA Type2	0.7017	미적용	CMA Type2	0.7148	미적용
		0.6931	n/4000		0.7107	n/4000



〈그림 7〉 아이템 기반의 예측에서 유의성 가중치에 따른 알고리즘과 유사도 가중치 별 MAE의 변화

웃 고객으로 어떠한 고객들이 선택되던지 상관없이 자신의 선호도가 동일하게 계산되기 때문에 그만큼 이웃과의 다각적인 관계를 반영하지 못하게 되는 것이다. CMA에서 자신의 선호도 뿐만 아니라 이웃의 선호도 또한 특정 고객과의 관계를 고려하여 계산되기 때문에 그 만큼 과중하게 계산된 이웃의 선호를 조정하게 되고 이는 선호도 예측 정확도에 반영된다.

추천시스템의 선호도 예측 정확도를 평가하는 MAE는 시스템 전체의 정확도를 평가하기 때문에 실제 개인별 정확도를 측정할 수는 없다. 본 연구에서는 실험집합의 전체 선호도 예측의 오차를 이용하는 시스템의 정확도 평가뿐만 아니라 개인별 선호도 평가의 오차를 이용하여 개인별 MAE를 계산하였고 이를 이용하여 통계적으로 각 알고리즘과 유사도 가중치에 따른 결과를 비교하였다. 본 연구의 실험에서 1개의 dataset에 대하여 사용자 기반의 예측 결과와 아이템 기반의 예측결과의 총 8가지의 결과를 얻게 되며 이 결과들 중 사용자 기반의 결과와 아이템 기반의 평균순위를 이용하여 알고리즘의 정확도에 대하여 통계적 분석을 하였다. 결과에서 사용자 기반의 예측과 아이템 기반의 예측에서 모두 상관계수를 이용한 CMA의 평균순위가 가장 낮아 가장 우수한 결과를 나타내고 있음을 알 수 있다.

고객 간의 유사관계를 나타내는 유사도 가중치로 사용된 피어슨 상관계수와 벡터 유사도는 두 고객이 공통으로 평가한 상품만을 이용하여 계산된다. 이 때 두 고객이 공통으로 평가한 상품의 개수가 적으면 두 고객의 관계가 올바르게 계산되지 않음을 알 수 있으며 이를 고려하여 유사도 가중치를 보정하는 유의성 가중치의 필요성이 필요한 것은 이미 알려져 있다. 그러나 이 유의성 가중치를 어떻게 설정할 것인지에 대하여는 일반화된 방법이 없음을 알 수 있다. 본 연구의 결과에서 기존 연구에서 제시된 유사도 가중치의 설정 값이었던 50 보다 그 이상의 값을 적용한 유의성 가중치에 의해 선호도 예측 정확도가 일정 수준까지 개선됨을 알 수 있으며 아이템 기반의 선호도 예측에서는 사용자 기반의 예측에서 보다 훨씬 많은 공통 평가 쌍의 개수가 예측의 정확도에 영향을 미치고 있음을 보여주고 있다. 그러므로 유의성 가중치 설정에 대한 좀 더 실증적인 연구가 필요하며 선호도 예측 정확도 향상을 일반화 시킬 수 있는 유의성 가중치를 설정하는 연구가 필요하다.

선호도 예측의 정확성을 향상시키고 추천시스템의 신뢰도를 향상시키기 위해 상품에 대한 선호도가 올바르게 반영되고 있는지에 대해 연구할 필요성이 있다. 이는 고의적으로나 악의적 의도로 선호도를 평가할 경우 이 선호도가 추천시스템에 대한 신뢰를 떨어뜨릴 수 있기 때문에 이를 제거할 수 있는 기법의 연구가 필요하다.

NBCFA와 CMA는 예측력이 우수한 알고리즘이지만 협업 필터링의 개념에서 제시된 바와 같이 이웃의 정보를 동시에 고려하게 된다. 이 때 이웃과의 선호도 유사정도를 나타내는 유사도 가중치는 본 논문에서 제시한 상관계수와 벡터 유사도가 사용되지만 이들 가중치는 고객과 상품의 수가 증가하고 규모가 확장되면 선호도 예측에 소요되는 시간이 크게 증가하게 된다. 실제 100K MovieLens dataset의 연구에서 계산된 공통 응답쌍의 개수와 Million dataset에서 계산되는 공통 응답쌍의 개수는 큰 차이를 보이고 있음을 알 수 있으며 이를 통해 추천시스템의 규모가 확장되면 신속한 추천에 큰 장애가 되고 있고 이를 해결하기 위한 연구가 필요하다.

참 고 문 헌

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

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

김재경, 안도현, 조윤호, "Development of a Personalized Recommendation Procedure Based on Data Mining Techniques for Internet Shopping Malls", 한국지능정보시스템학회논문지,

제9권, 제3호, 2003, pp. 177-191.

손재봉, 서용무, "협업 필터링 시스템에서 Degree of Match를 이용한 성능향상", Information Systems Review, 제8권, 제3호, 2006, pp. 139-154.

심장섭, "K-means 군집화와 순차 패턴 기법을 사용하는 VLDB 기반의 추천 시스템 설계", 충북대학교, 박사학위논문, 2005.

이희춘, 이석준, "대응평균 알고리즘을 이용한 협력적 필터링 추천시스템의 성능향상", 한국경영정보학회 2006 추계컨퍼런스, 2006, pp. 208-214.

한국인터넷진흥원, "2006년 상반기 정보화실태 조사", 2006.

Balabanovic, M., Y. Shoham, "Fab: content based, collaborative recommendation", *Communications of the ACM*, Vol.40, Issue 3, 1997, pp. 66-72.

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

Burke, R., "Semantic Ratings and Heuristic Similarity for Collaborative Filtering", *In Proceedings of AAAI Workshop on Knowledge based Electronic Markets 2000 (KBEM'00)*, Austin, TX. July, 2000.

Claypool, M., A. Gokhale, T. Miranda, P. Murnikov, D. Netes and M. Sartin, "Combining content based and collaborative filters in an online newspaper", *In Proceedings of ACM SIGIR Workshop on Recommender Systems: Algorithms and Evaluation*, University of California, Berkeley, Aug. 1999.

Deshpande, M., G. Karypis, "Item based top N recommendation algorithms", *ACM Transactions on Information Systems*, Vol.22 , Issue 1, 2004, pp. 143-177.

Goldberg, D., D. Nichols, B. M. Oki, D. Terry, "Using collaborative filtering to weave an information tapestry", *Communications of the ACM* , Vol.35, Issue 12, 1992, pp. 61-70.

Heckerman, D., D.C. Maxwell, C. Meek, R. Rounthwaite, C. Kadie, "Dependency Networks for Inference, Collaborative Filtering, and Data Visualization", *Journal of Machine Learning Research*, Vol.1, 2001, pp. 49-75.

Herlocker, J.L., J. Konstan, L.G. Terveen, J. Riedl, "Evaluating collaborative filtering recommender systems", *ACM Transactions on Information Systems*, Vol. 22, Issue 1, 2004, pp. 5-53.

Herlocker, J., J. Konstan, J. Riedl, "An Empirical Analysis of Design Choices in Neighborhood Based Collaborative Filtering Algorithms", *Information Retrieval*, Vol.5, No.4, 2002, pp. 287-310.

Hill, W.L., S.M. Rosenstein, G. Furnas, "Recommending and Evaluating Choices in A Virtual Community of use", *In Proceedings of the SIGCHI conference on Human factors in computing systems*, 1995, pp. 194-201.

Lee, H.C., "An Exploratory Study for Decreasing Error of Prediction Value of Recommended System on User Based", *Journal of the Korean Data & Information Society*, Vol.17, No.1, 2006, pp. 77-86.

Lin, W.S., A. Alvarez, C. Ruiz, "Efficient adaptive support association rule mining for recommender systems", *Data Mining and Knowledge Discovery*, Vol.6, 2002, pp. 83-105.

Mobasher, B., T.L. H. Dai, M. Nakagaxa, Y. Sun and I. Wiltshire, "Discovery of aggregate usage profiles for Web personalization",

In Proceedings of Workshop on Web Mining for E Commerce Challenge and Opportunities, 2000.

Resnick P. and H.R. Varian, "Recommender systems", *Communications of the ACM*, Vol.40, No.3, 1997, pp. 56-58.

Resnick, P., N. Iacovou, M. Suchak, P. Bergstorm, J. Riedl, "GroupLens: An Open Architecture for Collaborative Filtering of Netnews", *In Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work*, 1994, pp. 175-186.

Riedl, J., J. Konstan. "Word of Mouse: The Marketing Power of Collaborative Filtering", *New York: Warner Books*, 2002.

Sarwar, B.M., G. Karypis, J. A. Konstan, J. Riedl, "Item based collaborative filtering recommendation algorithms", *In Proceedings of Tenth International World Wide Web Conference*, 2001, pp. 285-295.

Sarwar, B.M., J. Konstan, A. Borchers, J. Herlocker, B. Miller, J. Riedl, "Using Filtering Agents to Improve Prediction Quality in the GroupLens Research Collaborative Filtering System", *In Proceedings of the 1998 Conference on Computer Supported Cooperative Work*, Nov. 1998.

Schafer, J. B., J. A. Konstan, J. Riedle, "E Commerce Recommendation Applications", *Journal of Data Mining and Knowledge Discovery*, Vol.5, 1/2, 2001, pp. 115-152

Shardanand, U.and P. Maes, "Social Information Filtering: Algorithms for Automating 'Word of Mouth'", *In Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems*, 1995, pp. 210-217.

Information Systems Review
Volume 9 Number 1
April 2007

A study on the Prediction Performance of the Correspondence Mean Algorithm in Collaborative Filtering Recommendation

Seok Jun Lee* · Hee Choon Lee**

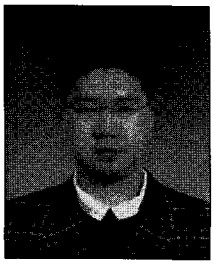
Abstract

The purpose of this study is to evaluate the performance of collaborative filtering recommender algorithms for better prediction accuracy of the customer's preference. The accuracy of customer's preference prediction is compared through the MAE of neighborhood based collaborative filtering algorithm and correspondence mean algorithm. It is analyzed by using MovieLens 1 Million dataset in order to experiment with the prediction accuracy of the algorithms. For similarity, weight used in both algorithms, commonly, Pearson's correlation coefficient and vector similarity which are used generally were utilized, and as a result of analysis, we show that the accuracy of the customer's preference prediction of correspondence mean algorithm is superior. Pearson's correlation coefficient and vector similarity used in two algorithms are calculated using the preference rating of two customers' co-rated movies, and it shows that similarity weight is overestimated, where the number of co-rated movies is small. Therefore, it is intended to increase the accuracy of customer's preference prediction through expanding the number of the existing co-rated movies.

Keywords: Recommender system, collaborative filtering, similarity weight, correspondence mean algorithm

* Adjunct Professor, Department of Business Administration, Sangji University
** Professor, Department of Computer Data & Information, Sangji University

◐ 저 자 소 개 ◑



이 석 준 (crco909@yahoo.co.kr)

상지대학교 산업공학과 및 산업환경대학원에서 석사를 마치고 일반대학원 경영학 박사학위를 취득하였으며 상지대학교 생산기술연구소 연구원을 역임하고 현재 상지대학교 경영학과 겸임교수로 재직 중이다. 관심분야는 전자상거래, 추천시스템, 데이터 마이닝 등이다.

이 희 춘 (choolee@sangji.ac.kr)

경희대학교 대학원 수학과에서 통계학으로 박사학위를 취득하고 동신대학교 통계학과 조교수를 역임, 강원대학교 컴퓨터과학과 박사과정 수료하고, 현재 상지대학교 컴퓨터 데이터 정보학과 교수로 재직 중이다. 관심분야는 정보검색, 전자상거래, 추천시스템 등이다.

논문접수일 : 2007년 02월 21일 **게재확정일 : 2007년 04월 17일**