# 협력적 추천시스템에서 유사도 가중치의 임계치 설정에 따른 선호도 예측 정확도 향상에 관한 연구

이석준
상지대학교 경상대학 경영학과
겸임교수
E-mail : crco909@yahoo.co.kr

전자상거래에서 거래되는 상품들에 대한 고객의 선호도를 사전에 파악하여 고객이 자신의 취향에 적합한 상품을 쉽게 찾도록 도와주고 전자상거래에 업체에 있어서는 목표고객의 설정을 자동적으로 처리할 수 있는 시스템이 추천시스템이다. 추천시스템은 고객과 업체 모두에게 이득을 가져올 수 있는 시스템으로 현재 많은 전자상거래 업체들이 적용 중에 있다. 본 연구는 전자상거래에서 널리 이용되고 있는 협력적 추천기법을 이용하여 고객 선호도 예측의 정확도를 향상시키기 위하여 고객들 간의 선호도 유사 정도를 나타내는 유사도 가중치에 일정 범위의 임계치를 설정하였다. 임계치의 설정에 따라 선호도 예측의 정확도가 향상되었으나 임계치의 설정 범위에 따라 고객 선호도를 예측할 수 있는 비율이 감소함을 알 수 있었으며 이에 따라 추천을 할 수 없는 고객이 발생할 수 있음을 알 수 있었다. 결과를 바탕으로 고객에 대한 추천과 예측의 정확도를 동시에 고려하는 임계치 설정에 대하여 더 많은 연구가 필요하다는 것을 알 수 있었다.

<색인어> 추천시스템, 유사도 가중치, 예측알고리즘, 임계치 설정

## 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 이용

논문제출일 : 2006. 11 게재확정일 : 2007. 3

자 중 최근 1년 이내에 인터넷을 이용하여 상품을 구매하거나 예약/예매 등 인터넷쇼핑을 한 적이 있는 인터넷쇼핑 이용률은 51.3%로 2005년 6월 48.2%에 비해 3.1%p 증가하였다(2006년 상반기 정보화실태조사-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 서비스를 제공하기 위한 전략을 생존의 필수적인 전략으로 인식하고 있다.

추천시스템은 전자상거래에서 고객의 입장과 기업의 입장을 모두 만족시킬 수 있는 시스템으로 받아들여지고 있다. 추천시스템은 고객의 취향이나 선호도를 미리 예측하여 고객이 선호하리라 생각되는 상품을 미리 추천하여 줌으로써 고객에게는 정보탐색비용의 절감과 기업에게는 판매상품에 대한 수요예측과 같은 자료와 목표고객의 설정 등과 같은 마케팅 전략에 다양하게 활용할 수 있다. 추천시스템은 많은 전자상거래 기업들이 도입하여 사용하고 있으며 대표적으로 Amazon.com은 추천시스템을 이용하여 다양한 마케팅 전략을 구사하고 있다. 따라서 추천시스템에서 정확한 상품 추천을 위한 추천 알고리즘의 개발과 알고리즘을 통한 예측치의 활용이 중요해지고 있다. 또한 전자상거래의 규모가 확대됨에 따라 전자상거래를 이용하는 고객의 수와 온라인으로 판매되는 상품의 종류와 수가 늘어나고 있다. 이에 따라 고객이 선호하리라 생각되는 상품에 대한 정확한 예측이 필요하게 되었으며 이를 위한 다양한 접근법이 제시되고 있다.

## 1. 연구 목적

본 연구는 추천시스템의 기법 중 협력적 필터링 기법의 추천 알고리즘인 이웃기반의 협력적 필터링 알고리즘(Neighborhood Based Collaborative Filtering)과 대응 평균 알고리즘(Correspondence Mean Algorithm)의 예측 결과를 개선하기 위하여 기존에 제시된 이웃 사용자와의 유사도 가중치에 대한 임계치(thresholder) 부여를 세분화하여 고객의 선호도 예측 정확도를 개선시키기 위한 방법들을 제시한다. 기존에 제시된 방법들은 MovieLens 100K dataset에 대한 분석이 이루어져 있으며 본 연구에서는 MovieLens 1million dataset을 대상으로 분석하여 유사도 가중치의 임계치 설정에 대한 영향을 분석하기로 한다.

# II. 이론적 연구

## 1. 추천시스템(Recommender System)

### 1.1. 추천시스템의 정의

전자상거래는 고객에게 다양한 종류의 제품이나 서비스를 서로 비교하여 상대적으로 품질이 양호하고 저가인 제품을 시간과 장소에 구애받지 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 개인화된 추천시스템(Personalized Recommender System)이라 부른다.

추천시스템은 제품, 서비스, 정보(일반적으로 통칭하여 제품)를 고객에게 추천하기 위해 다양한 응용분야에 이용되고 있다. 특히 전자상거래에서 추천시스템은 인터넷에 접목되어 고객과 제품 간의 관계(구매, 선호도, 그리고 카탈로그 검색 등과 같은 활동)에서 얻어지는 거래 데이터와 인터넷상에서 실시간으로 얻어지는 고객 행동에 대한 데이터를 분석하여 제품을 추천하기 위한 핵심적인 정보를 수집하게 된다. 이와 같이 고객과 제품 간의 관계 데이터는 고객의 인구통계학적 자료(고객의 신상명세)와 기업이 보유하고 있는 고객정보와 제품 정보로부터 향상된 정보를 얻게 된다. 이 모든 데이터들은 고객의 선호도 예측 모형과 알고리즘을 이용하여 고객이 원하는 상품을 추천하기 위한 추천시스템의 추천엔진에 입력되고 분석된다. 생성된 추천은 기업의 마케팅활동과 고객의 구매의사결정을 지원하기 위해 제공된다.

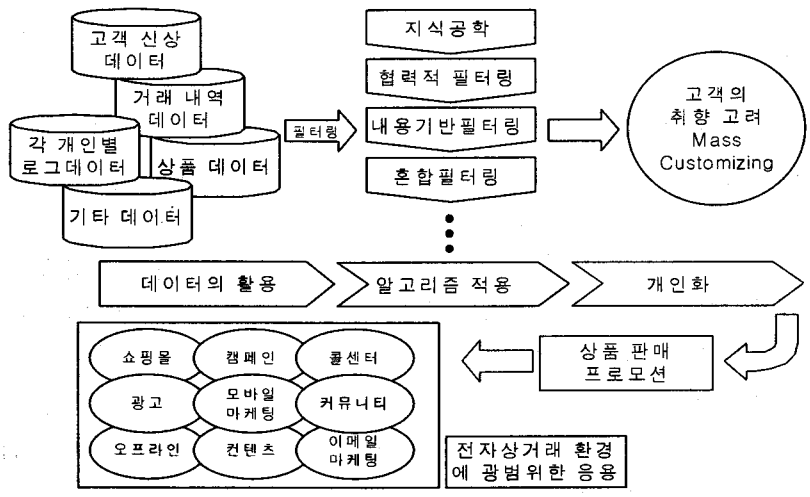
### 1.2. 추천시스템의 이점

Amazon.com, CDnow 등과 같은 전자상거래 기업들은 추천시스템을 성공적으로 활용하여 잠재 고객의 구매의도를 향상시키고 교차판매 효과를 증대시키며, 고객의 충성도를 향상시키는 효과를 얻을 수 있었다(Schafer, et. al., 2001). 전자상거래 기업에게 추천시스템은 고객의 구매 행위를 더 잘 이해하고 캠페인 활동의 효율성과 목적성, 교차판매와 같은 마케팅 활동을 전개하는데 유용한 도구로써 활용하고 있다. 효과적인 추천시스템의 활용은 전자상거래 사이트에서 직접적으로 판매 향상을 불러오는 것이라 할 수 있다. 추천시스템의 활용에 대한 정량적인 연구는 Konstan과 Riedl(2000)의 연구에서 NetPerceptions 추천시스템을 활용하여 평균 60% 향상된 교차판매의 효과를 얻었고 영국의 GUSpls

(www.gus.co.uk)사의 기법을 근간으로 하는 기존의 교차판매 기법보다 50% 향상된 성과를 거두었음을 보여주고 있다. 고객의 측면에서 추천 서비스는 인터넷에서 방대한 제품 중 구매하고자 하는 제품의 선택에 드는 탐색비용을 크게 줄일 수 있으며 특히, 인터넷에서의 전자상거래에서 더욱 큰 효과를 발휘한다(예; Netflix는 25,000편 이상의 영화를 제공하고, Amazon에서는 2백만 권 이상의 책을 제공하고, eBay의 경우 매일 수백만 건의 새로운 경매목록이 갱신된다.). 효과적인 추천은 고객의 쇼핑 경험을 향상시킬 수 있으며 고객의 수요증대와 충성도 향상을 가능케 할 것이다. 최근 연구에서는 Amazon의 음악 추천시스템에 의해 추천을 받은 20%의 제품이 구매되었다는 것을 보여주고 있다(Swearigen and Sinha 2002).

추천시스템을 이용하면 고객을 고정 시키는 효과가 있다. 일단 고객이 추천시스템의 서비스를 받아들이면 그에 따른 이득(검색비용의 절감과 성공적인 추천으로부터 얻어지는 뜻하지 않은 발견의 관점)은 다시 추천시스템의 추천 성능 향상에 더 큰 기여를 하게 된다. 영화 추천시스템의 예를 보면 "영화에 대한 더 많은 평가를 통해 더 나은 추천을 얻을 것임"으로 해석할 수 있다. 전자상거래의 특성상 추천시스템에 대한 고객의 기여는 전자상거래 사이트에서 구매로 이어지는 것으로 해석할 수 있다. 일단 고객이 특정 전자상거래 사이트의 추천서비스를 받아들이면 추천으로부터 얻어지는 이득을 얻기 시작하며 고객의 전환비용은 시간이 지남에 따라 증가하여 다른 경쟁 사이트로의 전환을 꺼리게 될 것이다. 이와 같은 고정 효과는 전자상거래 산업에서 고객의 충성도를 크게 향상시키며 추천시스템의 주요 전략적 가치를 제공하게 된다.

<그림 1> 추천시스템의 개요도(이희정, 2005)



전술하였듯이 고객에게 가치가 부여된 서비스를 제공하고 전자상거래 기업의 전략적 가치를 실현하는 것은 고품질의 추천을 생성할 수 있는 능력에 달려있다. 고품질의 추천시스템에 대한 요구는 추천기술을 연구하는 사업의 성장을 촉진하게 되었으며 많은 소프트웨어 기업들이 독자적인 추천기술을 제공하였다. NetPerception, Epiphany, Art Technology Group 등과 같은 상위 소프트웨어 공급사들은 2004년 12월까지 총 $6억 달러 이상의 시장 자본을 형성하였다(finance.yahoo.com). 한편 Amazon과 Netflix와 같이 초기에 추천시스템을 채택한 기업들도 자사에서 고도로 전문화된 추천기술을 개발하고 있다. 또한 최근에는 다양한 형태의 전자상거래 시스템에 결합된 독특한 추천시스템이 선보이고 있다(예: 여행산업을 위한 TripleHop Technologies사의 TripMatcher 시스템과 인터넷 음악방송을 위한 Yahoo의 LAUNCHcast 시스템).

## 2. 추천시스템의 분류

최초 추천시스템의 개념은 1992년 Goldberg(Goldberg, et. al., 1992)에 의해 주장되었고 다양한 기법들이 많은 응용영역의 추천시스템에 적용되었다. 초기 추천시스템은 Usenet news의 기사나 웹 페이지와 같은 대량의 문서에서 사용자들이 선호하는 내용을 여과하는 것을 도와주었다. 최근 전자상거래의 확산에 따라 추천시스템은 제품정보의 과잉에 따른 문제점을 극복하고 제품 선택에 대한 고객들의 능력을 향상시켜줄 수 있는 강력한 사업도구로 성장하고 있다.

추천시스템의 핵심은 다양한 형태의 입력 데이터를 이용하여 고객의 선호도에 부합하는 추천을 생성하기 위한 알고리즘에 있다. 추천에 대한 다양한 접근법은 인공지능(Artificial Intelligence)과 정보검색(Information Retrieval) 분야에서 개발되었다(Breese, et. al., 1998; Pazzani, 1999; Resnick and Varian, 1997). 이러한 대부분의 접근방법은 세 가지 형태의 입력 데이터를 취하며, 이 세 가지 입력 데이터의 형태는 제품의 특성, 고객의 특성, 그리고 소비자와 제품 간의 상호관계(구매와 평가 등)에 대한 데이터이다. 출력결과로는 상품에 대한 추천으로써 고객의 선호도에 대한 미래 예측 혹은 관찰되지 않은 고객과 제품 간의 상호관계이다. 입력 데이터의 형태에 따라 이러한 접근방법은 대략 내용기반(content-based), 인구통계학적 필터링(demographic filtering), 협력적 필터링(collaborative filtering), 그리고 혼합접근법(hybrid approaches)으로 나누어진다(Huang, et. al., 2004; Pazzani, 1999; Rensnick and Varian, 1997).

### 2.1 내용기반(Content-based) 추천기법

내용기반(content-based) 추천기법은 거래 상품의 속성을 분석하기 위해 정보검색 분야

의 기법을 이용하였다. 이웃(Neighborhood)의 기능(Claypool, et. al., 1999; Lieberman, 1995; Mladenic, 1996)과 분류화(Classification) 기법(Mooney and Roy, 2000; Mostafa, et. al., 1997)은 주로 상품의 특성에 대한 문자 내용을 분석하기 위해 적용되었다. 상품 추천은 이전에 고객이 구매하거나 경험한 상품의 특성에 대한 사용자 선호정도나 행동패턴과 같은 정보와 추천 대상이 되는 상품의 특성을 바탕으로 추천 대상 상품과 고객의 정보 간의 일치 정도를 기반으로 상품의 추천이 이루어진다. 일부 시스템들은 다방면의 매우 종합적인 상품에 대한 표현을 연구하였지만 내용기반의 접근법은 이전에 고객이 경험한 상품의 특성과 유사한 특성을 가진 상품에게만 추천이 이루어진다는 문제점이 있다. 내용기반 접근법의 근본적인 문제점은 상품을 추천하고자 하는 고객 자신의 경험이나 피드백만이 상품 추천에 이용된다는 것이다. 다시 말해, 내용기반 추천의 문제점은 고객의 선호도나 상품에 대한 흥미가 다른 고객들의 취향이나 선호도에 영향을 받음에도 불구하고 상품 추천을 위해 개별 고객 자신의 우선적인 경험만을 이용한다는 문제점이 있다(Balabanovic and Shoham, 1997).

## 2.2 인구통계학 기반(Demographic-based) 추천기법

인구통계학 기반의 추천기법은 내용기반의 추천기법과 유사하지만 이전에 상품을 구매하거나 경험한 고객들의 인구통계학적 특성을 근거로 상품 프로파일이나 정보를 생성한다. 생성된 상품 프로파일과 추천 대상이 되는 고객의 특성 사이의 일치성을 이용하여 상품을 추천한다는 점이 내용기반 추천기법과의 차이점이다. 인구통계학적 추천기법은 다른 고객들의 경험이나 취향을 상품 프로파일에 통합시켜서 유사한 고객들의 취향을 근거로 추천이 이루어진다는 점에서 내용기반의 추천기법 보다 좀 더 협력적인 성격을 가지고 있다. 이 접근법의 성과가 성공적이기 위해서는 고객에 대한 정보나 특성에 관한 자료가 매우 중요하다. 상품에 대한 정확한 고객의 취향이나 특성에 관한 정보는 일반적으로 구하기 어렵거나 구하는데 비용이 많이 든다. 인구통계학적 추천기법은 때로 협력적 필터링 접근법의 한 유형으로 구분되어지기도 하지만 고객의 특성에 따라 선호도가 명백히 구분되어지지 않고 고객이 경험한 상품을 기반으로 추천이 이루어진다는 점에서 협력적 필터링과 구분된다(Huang, 2005).

## 2.3 협력적 필터링(Collaborative filtering)

협력적 필터링 추천기법은 고객과 상품의 특성을 무시하고 고객-상품 간의 상호관계에 대한 데이터(예: 선호도 평가치)만을 이용하는 접근법을 취한다(Hill, et. al., 1995; Resnick, et. al., 1994; Shardanand and Maes, 1995). 협력적 필터링 추천기법은 고객과 상품 간의 상호관계나 피드백 데이터에 의존하기 때문에 고객과 상품 간의 상호관계는 이전의 상호

관계에 의해 명시적 자료가 아닌 암시적 자료로 구분된다. 상호관계 데이터만을 이용하는 추천의 간단한 예로 모든 고객들에게 가장 잘 알려진 상품을 추천하는 것을 들 수 있다. 또 다른 예로 상호관계 데이터를 분석하여 행렬 형태로 표현하고 분석하고자 하는 특성에 따라 추천을 하는 알고리즘을 예로 들 수 있다. 협력적 필터링은 가장 널리 이용되고 성공적인 추천 접근법으로 알려져 있으며 추천 알고리즘 연구의 근간을 이루고 있다(Breese, et. al., 1998; Resnick, et. al., 1994; Resnick, et. al., 1997).

다양한 범주의 협력적 필터링 알고리즘이 제안되었다. 대부분의 협력적 필터링 시스템은 사용자 기반(user-based)의 협력적 필터링시스템(Burke, 2000; Claypool, et. al., 1999; Mobasher, et. al., 2000; Nasraoui, et. al., 1999; Pazzani and Billsus, 1997; Sarwar, et. al., 1998)과 아이템 기반(item-based)의 협력적 필터링시스템(Deshpande and Karypis, 2004; Karypis, 2001; Sarwar, et. al., 2001)으로 나누어진다. 사용자 기반 협력적 필터링은 이웃 고객들에게서 얻어진 고객-상품 간의 상호관계 데이터를 이용한다. Pazzani의 연구에서 식당에 대한 추천을 얻고자 하는 고객과 식당을 이용한 고객들이 식당에 대한 선호정도의 표현을 바탕으로 유사 성향의 이웃 고객을 형성하여 추천이 이루어지도록 하였다(Pazzani, 1999). 아이템 기반의 협력적 추천기법은 상품들 간의 상호 유사관계를 이용하여 추천이 이루어지도록 한다. 또한 상품들 간의 관계는 고객-상품 상호관계를 바탕으로 이루어지고 이를 이용하여 추천이 이루어진다. 현재 전자상거래 추천시스템은 일반적으로 아이템 기반의 협력적 필터링 알고리즘을 이용한다(Schafer, et. al., 2001). 또 다른 협력적 추천시스템은 분석의 기초로 사용자와 아이템 중 단지 하나만을 이용하기 보다는 사용자와 아이템의 짝을 이루어 이용한다. 일부 시스템에서는 사용자 확인이 불가능하기 때문에 협력적 필터링을 거래 데이터에만 적용시킨다. 이웃 구성, 분류 알고리즘, 연관규칙 마이닝, 베이지안 네트워크, 그리고 군집화 모형과 같은 많은 데이터 분석 알고리즘이 협력적 필터링 문제에 적용되었다. 이웃은 코사인 벡터, 상관계수, 그리고 군집화 기법등과 같은 유사도 함수를 이용하여 구성되었다(Mobasher, et. al., 2000; Nasraoui, et al. 1999; Sarwar, et al. 2001). 일부 연구는 협력적 필터링 문제를 분류화(classification) 문제로 공식화 하였다(예를 들어, 고객-상품의 쌍으로 이루어진 데이터의 특성을 벡터로 정의하고 평가치를 부여하거나 거래 발생을 2진화하여 분류수준으로 결정하였다). 이후에 다양한 분류 알고리즘이 적용되었다. 연관규칙 마이닝은 거래기록에서 상품과 고객 간의 연관규칙에 대한 패턴을 끌어내는데 적용된다(Fu, et. al., 2000, Lin, et. al., 2002.). 표준 연관규칙 마이닝은 다양한 상품들(예를 들어 고객이 쇼핑에서 구매된 식료품)을 갖는 거래 내용에서 자주 거래가 발생하는 상품의 집단을 조사하고 주어진 상품 집단에서 관찰되는 상품을 조건부 확률을 바탕으로 연관규칙을 계산한다. 연관 규칙 추천 알고리즘은 협력적 필터링 문제를 다루기 위해 고객-상품 상호관계에 대한 특별한 구조를 처리하도록 표준 연관규칙 마이닝을 변형하게 된다.

연관규칙들의 집합은 상품들 사이의 선호도 종속성을 구체화시킨다. 베이지안 네트워크는 상품에 대한 고객의 경험들 간에 종속관계를 수학적으로 표현하는데 이 종속관계는 인과관계 혹은 상관관계를 반영하게 된다(Heckerman, et. al., 2001). 협력적 필터링에 대한 최근의 연구흐름은 고객과 상품의 2가지 요소로 구성된 데이터를 분석하기 위해 수학적 군집화 모형을 바탕으로 진행되고 있다(Hofmann and Puzicha, 1999; Kumar, et. al., 1998; Ungar and Foster, 1998). 이러한 모형은 사용자와 아이템들이 잠재적 형태 혹은 은닉된 군집에 속해 있다고 보고 군집화 모형의 변수를 측정하기 위해 기대치 최대화(Expectation Maximization) 방법과 같은 수학적 통계모형을 적용한다. 또한 K-means 군집화 알고리즘을 이용하여 거리개념의 군집을 생성한 후 군집간의 순차적 패턴을 발견하기 위한 방법도 연구되고 있다. 이러한 모형들은 주로 생성적 모형(generative models)로 불려지며 협력적 필터링 추천기법은 이와 같은 생성적 군집화 모형을 근거로 추천을 한다(심장섭, 2005).

### 2.4 혼합적 접근법(Hybrid Approach)

추천의 질을 향상시키기 위하여 내용기반 추천기법과 협력적 필터링의 접근법을 통합하기 위한 많은 노력들이 이루어졌으며 혼합적 접근법을 진행하게 되었다. 혼합적 접근법을 이용하는 추천 시스템은 대략 3가지 분류로 구분할 수 있다. 혼합적 시스템에서 첫 번째 분류는 내용기반 필터링과 협력적 필터링에서 얻어진 두 가지 개별적 추천 결과를 단순히 합치려는 시도이다(Claypool, et al., 1999). 두 번째 분류는 제품정보와 거래정보를 하나의 표현으로 통합하려는 표현 방식에 대한 통합의 시도이다. 이러한 시스템 중에 일부 시스템은 협력적 필터링에 기반을 두고 아이템에 대한 내용 정보를 사용자의 정보로 표시하거나 아이템의 정보를 기반으로 선호도를 부가하는 방법을 사용하였다(Pazzani, 1999; Sarwar, et. al., 1998; Singh, et. al., 2001). 예를 들어 Fab시스템(스탠포드 대학의 디지털 도서관 프로젝트의 일환)은 사용자에 의해 평가된 문서의 내용분석을 바탕으로 사용자 프로파일을 얻었으며 협력적 추천을 생성하기 위한 유사 사용자들을 구분하기 위해 사용자의 프로파일을 비교하였다(Balabanovic and Shoham, 1997). 다른 연구자들은 내용기반의 분석과 사용자의 정보를 아이템 정보에 부가하여 나타내기 위해 노력하였다(Goldberg et al., 1992). 세 번째 분류는 상이한 정보 자원을 통합하기 위한 포괄적인 모형을 설정하기 위한 시스템에 대한 연구이다(Basu, et. al., 1998; Condliff, et. al., 1999). 최근 Ansari 등(Ansari, et. al., 2000)의 연구에서는 아이템의 특성, 사용자의 특성, 그리고 전문가의 평가와 함수관계를 갖는 사용자의 선호도(rating)를 모형화 하기 위한 통계적 접근법을 적용하였다. 관측되지 않은 사용자의 선호도와 아이템의 이질적인 정보자원들은 이러한 추천 구조에서 사용되어야 한다고 밝히고 있다. 혼합적 접근 알고리즘의 포괄적 모형에서 중요한 분류는 생성적 군집 모형의 확장이다. 사용자와 아이템의 특징들뿐만 아니라 추가적인 관계 실체와 특

성들이 고품질의 추천을 위해 원칙적인 방법으로 생성적 군집화 모형에 통합되어야 한다.

### 2.5 지식공학적 접근법(Knowledge Engineering Approach)

지식공학적 접근법은 인간의 노력이나 학습적 방법을 통하여 제품을 선택하는데 사용자의 선호도에 영향을 미치는 요인들을 찾기 위한 접근법이다. 예를 들어 Expertise Recommender 시스템(McDonald and Ackerman, 2000)은 소프트웨어 공학전문가들이 기술지원 기록과 버전 변화 기록을 바탕으로 학습적인 방법을 이용하여 프로그램을 하도록 추천하였다. 학습적 방법은 5개월 간의 연구에서 얻어진 결과를 이용하여 만들어졌다. 메시지 필터링 시스템과 문서관리 시스템은 이와 같은 규칙 기반의 접근법을 널리 이용하여 왔다. 이러한 연구 분야의 초기 연구인 ISCREEN(Polloc, 1998)은 문자 메시지를 심사하기 위해 사용자의 지정된 규칙을 이용하였다. 규칙들은 각 사용자에 맞추어 개인화되었으며 규칙들은 주로 문자 메시지의 내용과 특성(발송자, 길이, 강조 등)을 바탕으로 만들어졌다. 지식공학적 접근법은 아이템의 특성, 사용자의 특성 그리고 특정 영역의 정보와 같은 넓은 영역을 바탕으로 추천을 목적으로 특정 사용자의 규칙을 만들어낼 수 있다. 이 접근법과 지금까지 살펴본 접근법과의 주요한 차이점은 추천을 위한 규칙을 만들어내기 위한 과정이 자동화된 것이 아니라 많은 사람의 노력이 필요하다는 점이다. 예를 들어 Burke는 지식기반 필터링 방법을 통해 얻어진 추천 결과를 협력적 필터링으로 다시 필터링 하는데 사용하는 혼합적 추천 구조를 제안하였다(Burke, 2000).

## 3. 협력적 필터링

### 3.1. 협력적 필터링의 개념

내용기반 추천기법은 정보의 과잉을 해소시키는데 효과적인 기법 중 하나이지만 다음과 같은 몇 가지 제한점이 있다(Shardanand and Maes, 1995).

● 추천 상품의 속성은 반드시 문자로 이루어져야 하며 멀티미디어 파일을 직접적으로 추천할 수는 없다.

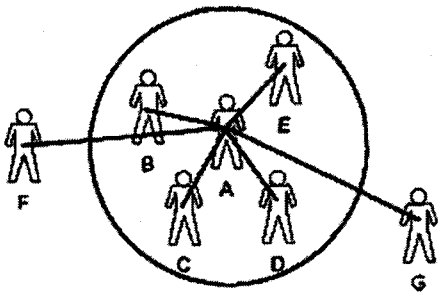
● 고객의 과거 경험만을 토대로 추천이 이루어지기 때문에 고객 본인의 선호도에만 특화된 상품만을 추천하고 뜻하지 않은 상품의 추천은 불가능하다.

● 상품의 품질이나 스타일, 유행, 혹은 상품에 대한 개인들의 견해 등을 반영한 추천을 할 수 없다.

협력적 필터링 추천기법은 내용기반 추천기법의 단점을 해결하기 위해 도입되었다. 협력

적 필터링 추천기법은 다른 고객들의 선호도를 이용하여 특정 고객의 상품 정보 과잉을 줄여주는데 목적을 두고 있으며 특정 고객과 유사한 선호도를 가지고 있는 다른 고객들의 평가를 기초로 상품을 추천한다. 이는 "구전 혹은 입소문"(Word of Mouth)을 정량화시키는 과정이다(Konstan, et. al., 1997; Shardanand and Maes, 1995).

<그림 2> 협력적 필터링의 개념(Bassam H. C., 2005)




협력적 필터링에서 이웃 고객들의 선호도 반영은 <그림 2>에서 설명된다. 예를 들어 어떤 상품에 대한 고객 A의 선호도를 예측하기 위해서 먼저 그 상품을 경험하거나 구매한 고객들을 이웃으로 선정하여야 한다. 그림에서 고객 A를 제외한 나머지 고객들이 어떤 상품에 대하여 선호도를 나타낸 고객이라 가정하면 고객 A의 이웃 고객은 고객 B, C, D, E, F, G가 되며 이들의 선호도 유사 정도를 계산하게 된다. 이때 선호도 유사 정도를 계량적으로 나타내기 위하여 유사도 가중치를 정의하여 사용한다. 구해진 유사도 가중치를 이용하여 각 고객들과의 유사 정도를 구분하는데 그림에서는 고객 A와 연결된 선의 길이를 유사 정도라 할 수 있다. 즉, 고객 F, G에 비하여 고객 B, C, D, E가 상대적으로 고객A 와 유사한 고객이라 할 수 있다. 어떤 상품에 대한 고객 A의 선호도를 예측하기 위하여 모든 이웃 고객들의 선호도를 이용할 수 있으며 그림에서와 같이 상대적 혹은 절대적인 기준에 따라 선호도가 유사한 고객들만을 이용하여 선호도를 예측할 수 있다. 절대적 방법은 유사도 가중치에 특정 임계값(threshold)을 설정하여 임계값 범위의 고객들의 유사도 가중치만을 이용하는 방법이고  상대적 방법은 고객 A와 유사 정도가 높은 고객 N명만을 선택하여 예측하는 방법이 될 수 있다.

내용기반 추천기법의 제한점을 해결하기 위해 도입된 협력적 필터링 기법도 다음과 같은 문제점을 가지고 있다.

● 초기 평가의 문제(first-rater): 어떤 상품에 대해 평가한 고객이 없을 경우 상품을 추천하지 못한다. 이는 추천의 기반이 되는 고객들의 선호도 평가치가 없기 때문이다. 또한

상품에 대한 선호도 평가치가 적을 경우 추천이 부정확하게 이루어질 수 있다. 이 문제는 협력적 필터링 추천기법을 도입한 시스템의 초기 운용에서 발생하는 문제점이기도 하다.

● 희소성의 문제(sparsity): 전자상거래에서 거래되는 상품의 개수는 고객이 관심을 가지고 찾고자 하는 양을 훨씬 넘어선다. 그래서 모든 상품에 대한 고객들의 선호도 평가치는 희소성을 가지게 된다. 밀도가 높은 데이터의 경우도 98~99%의 희소성을 가지기 때문에 추천에 충분한 자료를 가지지 못하여 예측이 정확하지 못할 수 있다.

● 모호집단(gray sheep): 중, 소규모의 전자상거래에서 고객집단과 상이한 선호도를 가진 고객의 경우는 정확한 추천을 받기 어렵다. 이는 추천을 위해 기반이 되는 이웃 고객들의 선호도와 잘 일치하지 않기 때문에 추천시스템이 정상적으로 기능을 발휘하더라도 정확한 추천을 받기 어렵다.

● 규모의 확장성(scalability): 전자상거래의 규모가 확장되고 이를 이용하는 고객의 수와 상품의 수가 증가하면 그만큼 추천시스템에 이용되는 데이터의 양이 증가하게 된다. 또한 동시 접속자 수가 폭주할 경우 예측 알고리즘의 계산 시간이 매우 길며 시스템의 리소스를 매우 많이 사용하게 된다.

이러한 문제점을 해결하기 위하여 많은 노력이 이루어지고 있으며 명시적 선호도 평가치가 아닌 암시적 선호도 평가치, 데이터의 분할, 차원 감소 혹은 정보 필터링 기법과의 통합에 대한 문제점을 극복하기 위한 연구들이 진행되고 있다(Good, et. al., 1999; Claypool, et. al., 2001; Zhang, 2002).

## 4. 추천 알고리즘

### 4.1 이웃 기반의 협력적 필터링 알고리즘 (Neighborhood-Based Collaborative Filtering Algorithm)

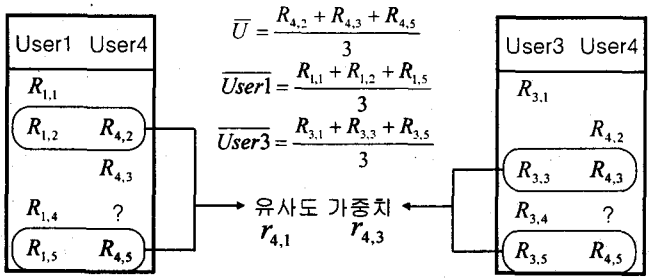
GroupLens에서 1994년에 처음으로 이웃 고객들의 선호도를 고려한 이웃 기반의 협력적 필터링 알고리즘(Neighborhood-Based Collaborative Filtering Algorithm)을 적용하여 유즈넷 뉴스(UseNet News) 그룹의 기사를 추천하였다. 유즈넷은 인터넷을 기반으로 형성된 토론 시스템으로 토론 주제에 따라 10여개의 뉴스그룹으로 구성되어 있으며 해당 뉴스그룹에 등록하면 토론 주제에 관심을 가지고 있는 수많은 가입자들과 의견을 교환하고 토론을 할 수 있는 시스템이다. GroupLens는 NBCFA을 이용하여 이웃 고객들의 선호도를 고려한 유즈넷 뉴스의 기사를 개인화시켜 고객이 원하는 뉴스 기사에 대한 추천을 자동적으로 예측하여 추천하였다(Resnick, et. al., 1994). 초기 GroupLens 시스템에서는 고객들 간의 선

호도의 유사 정도를 나타내기 위한 유사도 가중치(similarity weight)로 피어슨 상관계수(pearson's cprrelation coefficient)를 이용하였고 유즈넷 뉴스 그룹의 모든 고객들의 선호 기사에 대한 상관관계를 이용하였다. 특정 문서에 대하여 추천을 받을 고객의 선호도를 예측하기 위해 다음과 같은 이웃 사용자들의 평균과의 편차의 가중평균(weighted average of deviations from the neighbor's mean)을 이용하여 최종적인 선호도 예측치를 계산하였다.

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

여기에서 $\widehat{U_x}$는 추천을 받고자하는 고객 $u$가 특정 상품 $x$에 대해 어느 정도의 선호도를 보일 것인지에 대한 선호도 예측치이며 $n$은 추천을 받고자 하는 고객 $u$가 속해있는 그룹(예: UseNet News 그룹의 회원 혹은 전자상거래 사이트의 회원 등)의 이웃이 선호도를 표기한 상품의 수를 나타낸다. 또한 $r_{uj}$는 추천을 받고자하는 고객 $u$와 그룹 내의 이웃한 고객 $j$와의 선호도 유사 정도를 나타내는 유사도 가중치이다. $\overline{U}$는 추천을 받고자 하는 고객 $u$가 평가한 상품들의 선호도 평가치의 평균을 나타내며 $J_x$는 그룹 내의 이웃 고객 $j$가 평가한 특정 상품 $x$에 대한 선호도 평가치이다. $\overline{J}$는 고객 $u$가 추천을 받고자 하는 특정 상품 $x$에 대해 선호도를 평가한 이웃 고객 $j$의 선호도 평가치들의 평균이다. 이 때 $\overline{J}$는 고객 $u$에게 추천할 상품 $x$에 대한 선호도 평가치인 $J_x$를 제외한 사용자 $j$의 선호도 평가치들의 평균이다. Raters(선호도 평가자)는 test dataset에 포함되어 있는 상품에 대해 선호도를 평가한 고객들을 의미한다(Resnick, et. al., 1994).

<그림 3> NBCFA의 계산과정



<그림 3>에서 $\overline{User1}$는 User4와 이웃이 되는 고객의 평균으로 위 식의 $\bar{J}$에 해당하며 추천대상 고객의 선호도 예측을 하고자 하는 아이템에 해당하는 평가치를 제외하고 평균을 구하게 된다. 즉, $R_{1,4}$를 제외한 나머지 평가치들의 평균이 된다.

## 4.2 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 받고자하는 고객과의 선호도 유사 정도를 계량적으로 나타내어 이를 가중치로 적용할 필요성이 있다. 이때의 선호도 유사 정도를 나타내는 값을 유사도 가중치(similarity weight)라 하며 최초의 GroupLens 시스템에서는 피어슨 상관계수가 이용되었다(Resnick, et. al., 1994).

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

$$r_{uj} = \frac{\sum(U-\bar{U})(J-\bar{J})}{\sqrt{\sum(U-\bar{U})^2 \cdot \sum(J-\bar{J})^2}}, \quad -1 \le r_{uj} \le 1 \quad \cdots\cdots\cdots\cdots\cdots\cdots \quad (2)$$

$r_{uj}$는 추천을 받고자 하는 고객 $u$와 그룹 내의 이웃 고객 $j$가 선호도를 평가한 상품에 대한 피어슨 상관계수이며 이를 두 고객 간의 유사도 가중치로 이용한다. 여기서 $U$는 고객 $u$가 선호도를 평가한 상품의 선호도 평가치이고, $\bar{U}$는 고객 $u$가 선호도를 평가한 상품들의 선호도 평가치의 평균이며, $J$는 이웃 고객 $j$가 선호도를 평가한 상품의 선호도 평가치(rating)이고, $\bar{J}$는 이웃 고객 $j$가 선호도를 평가한 상품들의 선호도 평가치의 평균이다.

Breese 등(1998)은 피어슨 상관계수 외에 사용할 수 있는 유사도 가중치를 소개하고 평가하였다(Breese, et. al., 1998).

유사도 가중치는 피어슨 상관계수(pearson's correlation coefficient)뿐만 아니라 코사인 벡터를 이용한 벡터 유사도(vector similarity)를 이용하기도 하며 기본선호도(default voting), 역사용자 빈도(inverse user frequency), 사례확대(case amplification) 등의 다양한 유사도 가중치에 대하여 연구하였다(Breese, et. al., 1998).

벡터 유사도는 정보검색(information retrieval) 분야에서 두 문서간의 유사성을 구하기 위해 각 문서에서 단어의 출현 빈도를 벡터로 처리하여 두 빈도 벡터의 코사인 각도를 계산하여 유사성을 구하였다(Breese, et. al., 1998). 두 문서간의 유사성을 구하기 위한 코사인 벡터를 협력적 필터링에 적용하여 고객을 문서로 상품에 대한 선호도를 단어의 출현 빈도로 바꾸어 표현할 수 있다.

기본 선호도는 피어슨 상관계수를 확장한 것으로 기준이 되는 특정 고객과 그 고객과 선호도의 유사성을 구하게 되는 이웃 고객들이 공통으로 선호도를 평가한 평가치가 상대적으로 적을 경우에 사용된다. 기본 선호도는 어떤 고객도 선호도를 평가하지 않은 새로운 상품에 대하여 기본 선호도를 적용하여 추천이 가능하도록 할 수 있다. 그러나 기본 선호도를 설정할 경우 중립적인 선호도 평가치 혹은 다소 비선호의 선호도 평가치를 임의로 설정하여야 하기 때문에 추천시스템의 예측 정확도에 영향을 미칠 수 있다는 단점이 있다(정경용, 2005).

역사용자 빈도는 정보검색 분야에서 벡터 유사도를 이용하는 경우 문서에 포함된 단어의 수는 단어 빈도의 역수로 표현할 수 있다. 역사용자 빈도의 개념은 공통적으로 많이 발생하는 단어의 경우 문서를 분류하는데 유용한 역할을 할 수 없다는 판단에 따라 공통적으로 많이 발생하는 단어에 대하여 가중치를 줄이는 것이다. 협력적 필터링에서는 이러한 개념을 적용하여 일반적으로 사용자에 의해 많이 선호도가 평가되는 상품은 적게 선호도가 평가되는 아이템에 비해서 고객 간의 보여주는데 덜 기여한다고 판단하여 가중치를 줄이게 된다(정경용, 2005).

### 4.3 대응평균 알고리즘(Correspondence Mean Algorithm)

GroupLens에서 제시한 NBCFA을 수정한 대응평균 알고리즘(Correspondence Mean Algorithm)은 100K MovieLens dataset을 분석한 결과 NBCFA의 결과보다 향상된 예측력을 나타내었다(Lee, 2006). GroupLens의 분석결과에서는 선호도를 5점 척도로 구분한 자료에서 NBCFA의 MAE가 0.73의 자연적 장애물인 Magic Barrier에 도달한다고 밝히고 어떤 조정을 하여도 0.73이하로 잘 내려가지 않음을 지적하고 있다(Herlocker, et. al., 2004). 그러나 CMA의 경우 0.73 이하의 MAE를 나타내고 있다(Lee, 2006). NBCFA을 수정한 CMA를 살펴보면 다음과 같다. GroupLens에서 제시한 NBCFA에서 $\overline{U}$는 추천을 받고자 하는 고객이 나타낸 선호도 평가치 전체의 평균을 사용하고 있다. 이때의 $\overline{U}$는 고객 $u$의 선호도를 나타낸다. 그러나 특정 고객 $u$와 이웃 고객 $j$의 상관관계를 나타내는 가중치 $r_{uj}$(피어슨 상관계수)는 고객 $u$와 이웃 고객 $j$가 공통으로 평가한 상품의 선호도 평가치로만 구하게 된다. 여기서 $\overline{U}$를 고객 $u$의 선호도 평가치 전체의 평균을 이용하게 되면 고객 $u$의

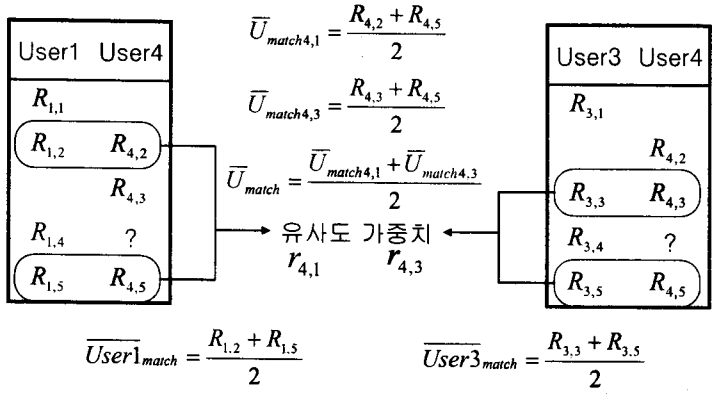
자신의 선호도가 과대평가되어 고객 $j$의 선호도를 충분히 반영하지 못하게 된다. 그래서 CMA는 고객 $u$와 이웃 고객 $j$가 공통으로 평가한 상품들의 선호도 평가치를 이용하여 고객 $u$의 선호도를 예측하기 때문에 이웃 고객 $j$의 선호도도 고객 $u$와 공통으로 평가한 선호도 평가치를 이용하였다. $\bar{J}$는 고객 $j$의 선호도를 나타내므로 고객 $j$가 평가한 아이템들만을 이용하면 이웃 고객 $j$의 선호도가 과대하게 평가된다. 그래서 고객 $j$의 선호도인 $\bar{J}$를 고객 $u$와 공통으로 선호도를 평가한 상품들의 선호도 평가치를 이용한 $\bar{J}_{match}$로 고객 $j$의 선호도로 정의하였다. CMA는 수식(3)과 같다.

$$U_x = \bar{U}_{match} + \frac{\sum_{J \in Raters} (J_x - \bar{J}_{match}) r_{uj}}{\sum_{J \in Raters} | r_{uj} |} \quad \cdots\cdots (3)$$

단, $\bar{U}_{match}$ : 고객 $u$와 이웃 고객 $j$가 공통으로 선호도를 평가한 상품들의 평가치를 이용한 고객 $u$의 선호도 평가치의 평균들의 평균

$\bar{J}_{match}$ : 고객 $u$와 이웃 고객 $j$가 공통으로 평가한 상품들의 평가치를 이용한 고객 $j$의 선호도 평가치의 평균

<그림 4> CMA의 계산과정




<그림 4>에서 $\bar{U}_{match4,1}$은 추천 대상 고객 User4와 이웃 고객 User1과의 유사도 가중치 계산에 이용되는 상품에 해당하는 평가치들의 평균으로 구해지며 또 다른 이웃, 즉,

이웃 고객 User3과도 동일한 방법으로 평균을 구한다. 이들 평균들의 평균을 $\overline{U}_{match}$로 정의하여 계산한다. $\overline{User1}_{match}$도 동일한 방법으로 유사도 가중치를 구할 때 이용되는 상품의 평가치로 계산하게 된다.

### 4.4 선호도 예측 정확도 평가척도

MAE(Mean Absolute Error)는 협력적 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, et. al., 1998; Herlocker, et. al., 1999; Shardanand and Maes, 1995).

$$MAE = \frac{1}{N}\sum_{j=1}^{N}\left| R_{uj} - \widehat{R_{uj}} \right| \quad \cdots\cdots (4)$$

여기서, N은 추천을 받을 모든 고객들에 대한 예측의 총 개수를 나타내며 $R_{uj}$는 실제 선호도 평가치이고 $\widehat{R_{uj}}$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.

## III. 연구방법론

### 1 실험데이터

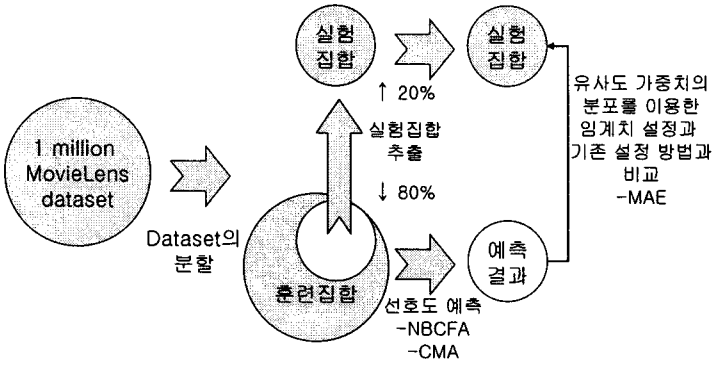
본 연구에서 사용된 dataset은 GroupLens의 Movielens dataset을 이용하여 실험을 하였다. GroupLens의 Movielens dataset은 미네소타 대학 Computer Science and Engineering 전공의 GroupLens research group에서 영화에 대한 사용자의 평가를 DB화 시켜 공개한 자료이다. MovieLens dataset은 100K dataset과 1million dataset의 2종류의 dataset이 있다. 100K dataset은 943명의 사용자와 1682편의 영화로 구성되어 있으며 영화에 대한 선호도 평가를 1~5점의 척도화 시킨 자료로 구성되어 있으며 총 100,000개의 선호도 평가치로 구성되어 있다. 본 연구에서는 GroupLens의 Movielens dataset 중 1million dataset을 이용

하여 분석하였다. 1million dataset은 6040명의 사용자와 3952편의 영화로 구성되어 있으며 6040명의 사용자들이 영화에 대해 평가한 선호도 평가치의 수는 1,000,209개 이다. GroupLens에서 제공하는 MovieLens dataset은 Ratings, Users, Movies의 3개의 파일로 구성되어 있으며 각각의 파일들은 다음과 같은 정보를 담고 있다.

## 2. 실험설계

본 연구에서는 MovieLens 1million dataset을 80%의 훈련집합(training set)과 20%의 실험집합(test set)으로 구분하여 유사도 가중치인 상관계수를 일정구간의 임계치를 설정하여 실험을 진행하였다. 알고리즘의 정확도를 평가는 훈련집합을 이용하여 고객들간의 유사도를 구하고 알고리즘을 적용하여 20%에 해당하는 실험집합의 평가치를 예측하여 평가척도를 이용하여 20%의 평가치와 해당 평가치에 대한 예측치와의 절대편차 평균을 이용하여 알고리즘의 정확도를 측정한다.

<그림 5> 실험의 전개도



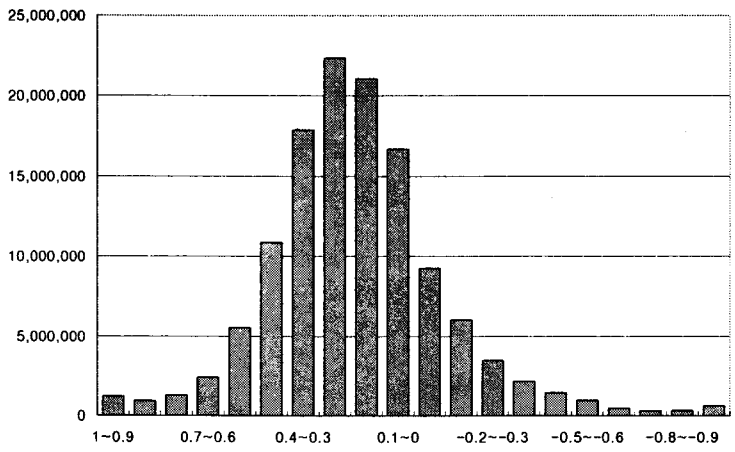

# IV. 실험

## 1. 유사도 가중치의 임계치 설정

본 연구에서 실험을 위하여 고객간의 유사도 가중치를 다음과 같이 임계치를 부여하여 구분하였다. herlocker 등의 연구에서 유사도 가중치의 임계치를 설정하기 위한 방법으로 일정 수준이상의 상관계수, 즉 양의 상관계수와 응답 쌍, 즉 상관계수를 구하기 위한 두 고

객간에 공통으로 선호도가 평가된 상품의 평가치들의 개수를 제한하는 방법을 제시하였으며 양의 상관계수를 이용하였을 경우가 전체 상관계수를 이용하는 것 보다 예측의 정확도가 높다는 것을 보여주었다. 본 연구에서는 임계치를 설정하지 않은 전체 상관계수를 이용하였을 경우와 herlocker의 연구에서와 같이 양의 상관계수만을 이용하였을 경우 그리고 상관계수의 분포, 즉 유사도 가중치의 분포를 이용하여 통계적 방법인 ± 1σ, ± 2σ, ± 3σ의 범위에 존재하는 상관계수일 경우 선호도 예측에 참여시키도록 하였으며 이는 과도하게 높은 상관계수와 혹은 반대의 경우의 상관계수는 전체의 상관계수에서 노이즈에 해당할 것으로 가정하고 임계치를 설정하였다. 또한 유사도 가중치 분포의 평균 이상이 되는 상관계수만을 이용하는 방법과 예측비율(coverage)을 손상시키지 않는 범위의 임계치인 양의 상관계수 중 0.3 이상인 경우로 나누어 결과를 비교하였다. 훈련집합의 유사도 가중치는 <그림 6>과 같은 분포를 나타내고 있으며 외형적으로 정규분포에 근사함을 알 수 있으며 유사도 가중치의 평균은 0.1808이고 표준편차는 0.0789로 계산되었으며 훈련집합에서 고객간의 관계를 나타내는 유사도 가중치의 총 개수는 125,686,032개로 분석되었다.

<그림 6> 훈련집합의 유사도 가중치 분포




## 2. 실험결과

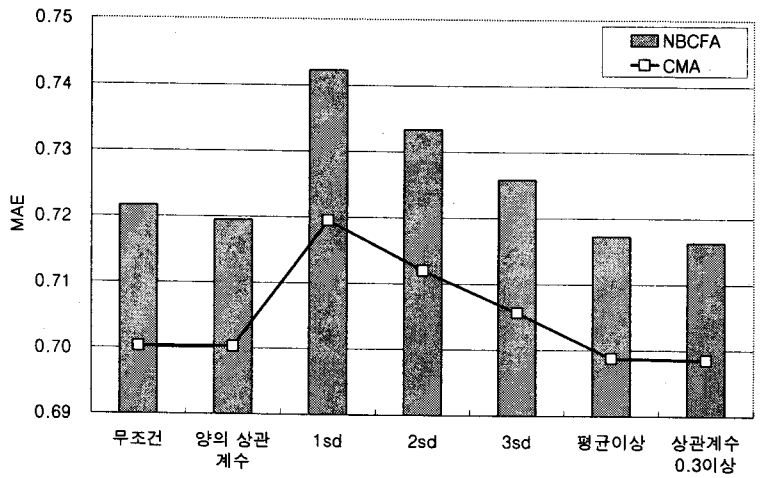
<표 1>은 연구에서 설정한 임계치와 그에 따른 알고리즘 별 MAE의 결과표이고 <그림 7>은 결과도이다. 결과에서 무조건은 유사도 가중치에 임계치를 부여하지 않은 결과를 나타내며 양의 상관계수는 유사도 가중치인 상관계수에서 음의 상관계수를 제외한 양의

상관계수를 가지는 유사도 가중치만을 예측 알고리즘에 활용하였다. ±1σ, ±2σ, ±3σ는 유사도 가중치의 평균에 각각 표준편차를 이용한 범위를 임계치로 설정하였을 경우의 결과이며 평균이상은 유사도 가중치의 평균인 0.1808이상의 유사도 가중치만을 활용한 결과이다. 또한 상관계수 0.3이상은 유사도 가중치의 결과가 0.3이상일 경우만을 예측에 활용하였을 경우의 결과이다.

<표 1> 임계치 설정에 따른 각 알고리즘의 MAE

| 구분 | NBCFA | CMA |
|---|---|---|
| 무조건 | 0.72167 | 0.70029 |
| 양의 상관계수 | 0.71952 | 0.70027 |
| ±1σ | 0.74229 | 0.71951 |
| ±2σ | 0.73340 | 0.71200 |
| ±3σ | 0.72587 | 0.70565 |
| 평균이상 | 0.71729 | 0.69875 |
| 상관계수 0.3이상 | 0.71632 | 0.69850 |

<그림 7> 임계치 설정에 따른 각 알고리즘의 결과도




다음 <표 2>와 <표 3>은 NBCFA와 CMA에 따른 개인별 MAE의 평균순위 비교검증으로 <표 2>에서 NBCFA를 이용하였을 경우 평균이상의 임계치를 부여하였을 때 통계적으로 가장 순위가 우수한 것으로 분석되었으며 <표 3>에서 CMA를 이용하였을 경우도 <표 2>의 유사한 결과를 얻음을 알 수 있다.

<표 2> 임계치 설정에 따른 NBCFA의 개인별 MAE의 프리드만 평균순위 검증 결과

| 구분 | N | 평균 | 표준편차 | 평균순위 | $\chi^2$ | 유의확률 |
|---|---|---|---|---|---|---|
| 무조건 | 6032 | 0.747 | 0.232 | 4.086 | 2332.818 | 0.000** |
| 양의 상관계수 | | 0.740 | 0.231 | 3.643 | | |
| ±1σ | | 0.756 | 0.237 | 4.902 | | |
| ±2σ | | 0.750 | 0.235 | 4.503 | | |
| ±3σ | | 0.745 | 0.234 | 3.958 | | |
| 평균이상 | | 0.739 | 0.230 | 3.404 | | |
| 상관계수 0.3이상 | | 0.739 | 0.230 | 3.504 | | |

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

<표 3> 임계치 설정에 따른 CMA의 개인별 MAE의 프리드만 평균순위 검증 결과

| 구분 | N | 평균 | 표준편차 | 평균순위 | 카이제곱 | 유의확률 |
|---|---|---|---|---|---|---|
| 무조건 | 6032 | 0.729 | 0.225 | 3.905 | 2077.08 | 0.000** |
| 양의 상관계수 | | 0.723 | 0.221 | 3.637 | | |
| ±1σ | | 0.739 | 0.228 | 4.883 | | |
| ±2σ | | 0.733 | 0.226 | 4.508 | | |
| ±3σ | | 0.728 | 0.224 | 3.988 | | |
| 평균이상 | | 0.722 | 0.221 | 3.489 | | |
| 상관계수 0.3이상 | | 0.723 | 0.220 | 3.591 | | |

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

## V. 시사점 및 결론

본 연구의 결과를 통하여 유사도 가중치에 임계치를 설정하여 고객 선호도를 예측하는 방법이 협력적 추천시스템의 성과를 향상시킬 수 있다는 것을 알 수 있었다. 그러나 과도한 임계치의 설정은 예측 결과를 개선시키지 못하며 또한 고객에 대한 예측 비율을 감소시킬 수 있음을 알 수 있었다. 특히 유사도 가중치의 평균과 표준편차를 이용하는 방법은 전체 실험집합에 구성되어 있는 영화의 평가치를 모두 예측하지 못하는 것을 알 수 있다. ±3σ의 경우 예측치의 개수가 200,001인데 반하여 ±1σ의 경우 199,855개의 예측 만 할 수 있었으며 유사도 가중치인 상관계수가 0.3이상일 경우 199,960개의 예측 만 할 수 있었다. 이는 과도한 임계치의 설정은 고객에 대한 충분한 선호도 예측치를 만들지 못하기 때문에 고객들의 만족도를 떨어뜨릴 가능성이 있다.

본 연구에서는 이러한 예측 비율에 대한 연구는 언급하지 않았지만 추천시스템의 성능을 높이기 위한 임계치 설정방법은 예측 정확도의 향상과 더불어 예측 비율 양쪽을 고려하여 분석할 필요성이 있으며 차후에 심도 깊은 연구가 필요하다.

# 참 고 문 헌

손재봉, 서용무, 2006. "협업 필터링 시스템에서 Degree of Match를 이용한 성능향상", Information Systems Review, 8-2, 139-154.

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

이희정, "데이터마이닝을 이용한 eCRM 환경에서의 추천시스템", 부산대학교, 석사학위논문, 2005

정경용, "혼합 필터링과 연관 이웃 마이닝을 이용한 개인화 아이템 추천 기법", 인하대학교, 박사학위논문, 2005

Adomavicius, G. and A. Tuzhilin. 2001. "Using Data Mining Methods to Build Customer Profiles," IEEE Computer, 34-2, 74-82.

Al Mamunur Rashid, Shyong (Tony) K. Lam, George Karypis, and John Riedl. "ClustKNN: A Highly Scalable Hybrid Model- & Memory-Based CF Algorithm" WEBKDD 2006, August 20, 2006, Philadelphia, Pennsylvania, USA

Ansari, A., S. Essegaier and R. Kohli. 2000. "Internet Recommendation Systems," Journal of Marketing Research, 363-375.

Bassam H. C., 2005, "Use of Discrete Choice Models with Recommender Systems", Doctoral Thesis, MIT.

Basu, C., H. Hirsh and W. Cohen. 1998. "Recommendation as Classification: Using Social and Content-based Information In Recommendation," In the Proceedings of Fifteenth National Conference on Artificial Intelligence, pp. 714-720, Madison, Wisconsin, July 1998.

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, pp. 43-52, Madison, Wisconsin, July 1998.

Burke, R. 2000. "Semantic Ratings and Heuristic Similarity for Collaborative Filtering," In the 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. 1999.

"Combining content-based and collaborative filters in an online newspaper," In the Proceedings of ACM SIGIR Workshop on Recommender Systems: Algorithms and Evaluation. University of California, Berkeley, Aug. 1999.

Condliff, M. K., D. D. Lewis, D. Madigan and C. Posse. 1999. "Bayesian Mixed-effects Models for Recommender Systems," In the Proceedings of ACM SIGIR Workshop on Recommender Systems: Algorithms and Evaluation. University of California, Berkeley, Aug. 1999.

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

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

H. C. Lee. 2006. "Improved Algorithm for User Based Recommender System", Journal of the Korean Data & Information Science Society, Vol 17. No. 3, 717-726.

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

Hofmann, T. and J. Puzicha. 1999. "Latent Semantic models for collaborative filtering," In the Proceedings of International Joint Conference in Artificial Intelligence, Stockholm, 688-693.

Konstan, H. J., Borchers, J. A., Reidl, J. 1999. "An Algorithmic Framework for Performing Collaborative Filtering", In Proceedings of the 1999 Conference on Research and Development in Information Retrieval.

Konstan H. J.,Reidl J., 2000. "Tutorial notes: Recommender systems in e-commerce," In the Proceedings of ACM E-Commerce 2000 Conference.

Liberman, H. 1995. "Letizia: An agent thatassists Web browsing," In the Proceedings of International Joint Conference on Artificial Intelligence, Montreal, 924-929.

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

Mladenic, D. 1996. "Personal Web Watcher: Implementation and design," Technical Report, IJ

Mobasher, B., T. L. H. Dai, M. Nakagaxa, Y. Sun and I. Wiltshire. 2000. "Discovery of

aggregate usage profiles for Web personalization," In the Proceedings of Workshop on Web Mining for E-Commerce-Challenge and Opportunities.

Nasraoui, o., H. Frigui, A. Joshi and R. Krishnapuram. 1999. "Mining Web access logs using relational competitive Fuzzy clustering," In the Proceedings of Eighth International Fuzzy Systems Association World Congress-IFSA 99.

Pazzani M. and D. Billsus. 1997. "Learning and revising user profiles: the identification of interesting Web sites," Machine Learning, Vol. 27, No. 3, 313-331.

Pazzani, M. 1999. "A framework for collaborative, content-based and demographic filtering," Artificial Intelligence Review, Vol. 13, No 5, 393-408.

Popescul, A., L. H. Ungar.D. M. Pennock and S. Lavrence. 2001. "Probabilistic models for unfied collaborative and content-based recommendation in sparse-data environments," In the Proceedings of 17'th Conference on Uncertainty in Artificial Intelligence(UAI 2001).

Resnick, P., Iacovou, N., Suchak, M., Bergstorm, P., 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.

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

Schafer, J. B., Konstan J. A., Riedle, J. 1999. "Recommender systems in e-commerce", ACM Conference on Electronic Commerce, 158-166.

Schafer, J. B., Konstan J. A., Riedle, J. 2001. "E-Commerce Recommendation Applications", Data Mining and Knowledge Discovery

Schwab, I., W. Phol and I. Koychev. 2000. "Learning to recommend from positive evidence," In the Proceedings of International Conference on Intelligent User Interfaces, 241-247.K. In the poroceeding of Designing Interactive System, 2002.

Swearingen K. and R. Sinha, 2002, "Interaction design for recommender systems",

Ungar, L. H. and D. P. Foster. 1998. "A formal statistical approach to collaborative filtering," In the Proceedings of CONALD'98.

# A Study on the Improvement of Prediction Accuracy of Collaborative Recommender System under the Effect of Similarity Weight Threshold

**Seok-Jun Lee**

## Abstract

Recommender system helps customers to find easily items and helps the e-biz companies to set easily their target customer by automated recommending process. Recommender systems are being adopted by several e-biz companies and from these systems, both of customers and companies take some benefits. This study sets several thresholds to the similarity weight, which indicates a degree of similarity of two customers' preference, to improve the performance of prediction accuracy. According to the threshold, the accuracy of prediction is being improved but some threshold setting shows the reduction of the prediction rate, which is the coverage. This coverage reduction has male effect on the prediction accuracy of customers, so more study on the prediction accuracy of recommender system and to maximize the coverage are needed.

Key Words : Recommender system, similarity weight, prediction algorithm, threshol.