IE Interfaces
Vol. 19, No. 1, pp. 70-77, March 2006.

속성추출을 이용한 협동적 추천시스템의 성능 향상

유상종 · 권영식†

동국대학교 산업시스템공학과

Performance Improvement of a Collaborative Recommendation System using Feature Selection

SangJong Yoo · Young S. Kwon

Industrial and Systems Engineering, Dongguk University, Seoul, 100-715

One of the problems in developing a collaborative recommendation system is the scalability. To alleviate the scalability problem efficiently, enhancing the performance of the recommendation system, we propose a new recommendation system using feature selection. In our experiments, the proposed system using about a third of all features shows the comparable performances when compared with using all features in light of precision, recall and number of computations, as the number of users and products increases.

Keywords: collaborative recommendation systems, feature selection, data mining

1. 서 론

1.1 연구 배경

우리나라의 초고속 인터넷 가입자수는 1998년 이후 연평균 160%의 성장을 보여 2005년 11월 현재 약 1,200만 명에 달하고 있으며, 인터넷 사용자는 약 3,300만 명으로 전 국민의 약 72%가 인터넷을 이용하고 있다(Ministry of Information and Communication, 2005).

이와 같은 인터넷 인프라를 기반으로 우리나라의 전자상거래 규모는 2000년 58조 원에서 2004년 300조 원으로 4년만에 약 400%의 급성장을 보이고 있다. 또한 전 세계의 전자상거래 규모도 2001년 3,300억 달러에서 2004년에는 약 3조 달러에 이르고 있다(Korea Institute For Electronic Commerce, 2005).

이와 같이 전자상거래 시장규모가 급성장함에 따라 취급되는 상품의 수와 종류도 크게 증가되어, 온라인 쇼핑몰은 소비자들의 구매패턴이나 구매이력들을 분석하여 소비자들이 원하는 상품을 편하게 찾고 구매추천을 할 수 있는 추천시스템을 적극적으로 활용하고 있다.

추천시스템은 인구통계학적 자료를 이용한 추천, 내용기반 추천, 항목기반 추천, 협동적 추천 등의 방법이 있으며 본 연구는 많이 이용되는 협동적 추천방법을 연구대상으로 한다.

협동적 추천방법은 사용자의 구매데이터를 기반으로 유사도가 높은 그룹을 형성하고 그룹에 속한 사용자의 구매데이터를 이용하여 특정 사용자를 위한 상품추천을 한다(Billsus and Michael, 1998).

협동적 추천시스템은 사용자 간의 구매 유사성을 찾아내기 위해 구매데이터를 이용하여 유사도를 계산한다. 이러한 방법은 시스템의 운영 초기에는 자료가 많지 않기 때문에 적절한 유사도를 구하기 어려운 반면, 시스템이 운영됨에 따라 누적되는 사용자 수와 엄청난 구매자료로 인하여 유사도계산에 많은 시간이 걸려 추천속도가 크게 저하되는 확장성(scalability)이 문제점이었다.

1.2 연구 목적

온라인 쇼핑몰의 추천시스템은 고객에게 실시간 또는 주기적으로 제품에 대한 새로운 정보를 제공하기 때문에 추천을

†연락저자 : 권영식 교수, 100-715 서울시 중구 필동3가 26번지 동국대학교 산업시스템공학부, Fax : 02-2260-3378,
E-mail : yskwon@dongguk.edu
2003년 12월 접수, 1회 수정 후 2005년 12월 게재 확정.

위해서 걸리는 시간이 매우 민감한 문제이다. 본 연구의 목적은 구매자료가 누적됨에 따라 증가되는 추천시간을 줄이기 위해 사용자들의 구매데이터로부터 중요한 속성을 추출함으로써 사용자 간의 유사도 계산시간을 줄여 확장성(scalability)문제를 해결하는 것이다.

2. 추천시스템에 대한 기존 연구

추천시스템은 인구통계학적 데이터를 이용한 추천, 내용기반 추천, 항목기반 추천, 협동적 추천시스템 등이 있다.

인구통계학적 자료를 이용한 추천은 사용자의 성별, 나이, 직업 등과 같은 인구통계학적 요소에 의해 사용자 유형별 특징을 분석하여 상품을 추천하는 방법이다. 이 기법은 단순한 정보 필터링 기법으로 구현방법이 간단하고 사용자로부터 피드백 정보가 없어도 추천이 가능하며 초기의 추천시스템에 활용 되어 왔다.

내용기반 추천기법은 개인이 입력한 정보와 상품에 관련된 텍스트 정보를 이용하여 추천하는 방법이다. 이 방법은 논리연산자로 결합된 검색어를 이용하여 정보 필터링을 하며 필터링에 사용된 프로파일은 자동적으로 업데이트된다. 그러나 정보 필터링 과정에서 개인과 상품의 프로파일 정보만 이용함으로써 필터정보가 제한적이고, 효과적인 추천을 위해서는 상품에 대한 상세한 속성정보가 필요하다(Balabanovi and Shoham, 1997).

내용기반 추천기법은 사용자 프로파일을 통해 과거 구매나 추천결과를 쉽게 반영할 수 있으며 추천속도가 빠르다. 그러나 상품에 대한 텍스트 데이터의 정확도를 판단하기 어렵고 상품과 사용자가 많은 경우에 효율성이 떨어진다(Claypool *et al*. 1999).

항목기반추천 기법은 상품 간의 유사성을 이용하여 상품을 추천하는 방식이다. 상품 간의 관계를 기반으로 하나의 상품에 대한 구매결과로부터 다른 상품에 대한 구매를 이끌어 내는 방법으로 코사인 계수와 조건부 확률을 이용해서 상품간 유사도를 계산하여 사용자의 바구니에 들어 있는 상품과 유사도가 높은 후보상품을 추천한다(Sarwar *et al*. 2001; Karypis, 2000). 이 방식은 상품에 대한 평가결과가 적은 초기 시스템에서는 사용자 간의 유사성을 찾는 협동적 필터링 기법보다 유용하지만, 사용자가 평가한 상품과 유사한 상품이 많고 실시간으로 운영되는 환경에서 모든 상품 간 유사도를 실시간으로 계산하기 어렵다는 문제점을 가지고 있다.

협동적 추천기법은 타깃(target) 사용자와 유사한 선호도를 가지는 다른 사용자의 상품에 대한 평가를 이용하여 타깃 사용자에게 적절한 상품을 추천하는 방식이다. 먼저 타깃 사용자의 과거 평가결과를 이용하여 가장 유사한 사용자 그룹을 선택하고, 이 유사그룹의 사용자가 이미 평가한 상품 중에서 타깃 사용자가 평가하지 않은 상품에 대한 평가값을 예측하여 추천한다. 이 방식은 사용자 기반의 정보 필터링 방식으로, 고객 개인별 추천이 가능하며 예측 정확성이 높다.

본 연구에서는 협동적 추천기법에 초점을 맞추고자 한다.

2.1 협동적 추천에 관한 기존 연구

협동적 추천에 관한 연구는 추천을 위해 사용자 간 유사도를 계산하는 부분과 시스템 운영 초기의 추천에 사용되는 데이터의 양이 적어서 생기는 '희소성'의 문제를 해결하기 위한 연구 등 크게 두 방향으로 연구가 진행되고 있다.

Breese *et al*.(1998)는 피어슨 상관계수와 벡터 유사도(vector similarity) 방법을 사용하고 기본값, 역사용자 빈도수(inverse user frequency), 사례확대(case amplification)의 방법을 이용하여 정확도와 유효범위를 향상시키는 것에 관한 연구와 기존의 확률적 방법인 베이지안(Bayesian) 방식의 모델 기반(model-based) 방법을 협동적 방법에 응용하였다.

Helocker *et al*.(1999)은 다양한 방식의 유사도 계산과 여러 가지 방식의 유사도 가중치 실험을 하였다. 유사도 계산에는 피어슨 상관계수, 스피어만 상관계수, 벡터 유사도를 이용하고, 선호도 값을 구하는 방법으로는 평균등급(average rating), 유사 사용자의 상품 선호도 가중치 합(deviation from mean), z 평균점수(z score average)방법을 이용하여 실험을 했다. 실험결과로서, 유사도를 구할 때 평가하는 선호도 값의 범위가 연속적인 경우에는 피어슨 상관계수를 이용하는 것이 높은 정확도를 나타냈고, 선호도 값을 구할 때는 전체적으로 유사 사용자의 상품 선호도 가중치 합을 이용하는 것이 높은 정확도를 나타냈다.

Billsus *et al*.(1998)은 상품추천에 소요되는 시간을 단축하기 위하여 특이행렬분해(singular value decomposition) 방법을 이용하여 상품의 차원을 줄임으로써 시간을 단축하고 유효범위를 개선하고자 하였다.

Kim(2001)은 DEC 사의 EachMovie 데이터 자료로 희소성 문제를 해결하기 위해서 특이행렬분해를 적용한 연구를 했다.

협동적 추천방법을 이용하여 개발된 추천시스템에는 GroupLens(Konstan *et al*., 1997), Ringo(Shardanand, 1995), Fab(Balabanovi and Shoham, 1997), Siteseer(Rucker and Polanco, 1997) 등이 있다. Amazon.com, CDNow, MovieFinder.com, Launch.com 등이 협동적 추천기법에 의한 추천시스템을 이용하고 있다(Schafer *et al*. 1999).

(1) GroupLens(Konstan *et al*., 1997)

GroupLens는 흥미 있는 Usenet 뉴스기사를 찾기 위해서 협동적 추천방식을 사용하였다. 이 시스템은 각 사용자 평가데이터와 추천데이터를 가지고 유사 사용자를 찾기 위해서 피어슨 상관계수를 적용했다. 식 (1)의 피어슨 상관계수를 통해서 사용자 a와 사용자 k 사이의 유사도의 정도를 결정하고 이렇게 계산된 유사도를 이용하여 식 (2)와 같이 사용자 a의 정보 i에 대한 선호도를 예측하게 된다. 유사도가 가중치로 이용되므로

유사도가 높은 사용자의 선호도 예측에 더 많이 반영된다.

$$w_{a,k} = \frac{\sum_j (r_{a,j} - \overline{r_a})(r_{k,j} - \overline{r_k})}{\sqrt{\sum_j (r_{a,j} - \overline{r_a})^2 \times \sum_j (r_{k,j} - \overline{r_k})^2}} \quad (1)$$

$$p_{a,i} = \overline{r_a} + \frac{\sum_k w_{a,k} \times (r_{k,j} - \overline{r_k})}{\sum_k w_{a,k}} \quad (2)$$

j :사용자a, 사용자k 가 선호도를 가지는 정보,
$r_{a,j}$: 사용자a의 정보 j에 대한 선호도,
$r_{k,j}$: 사용자k의 정보 j에 대한 선호도,
$\overline{r_a}$: 사용자a의 전체 정보에 대한 평균 선호도,
$\overline{r_k}$: 사용자k의 전체 정보에 대한 평균 선호도

(2) Ringo(Shardanand, 1995)

Ringo는 음악가에 대한 평가를 웹 또는 전자우편을 통해서 받아들여 사용자 프로파일을 구성하는 음악 추천시스템이다. 사용자들이 자신들이 좋아하거나 싫어하는 음악가에 대한 프로파일을 작성하기 때문에, 사용자 간 유사한 성향을 가진 다른 사용자를 찾을 수 있다. 예를 들면, 사용자1이 음악가 M1과 M2를 선호하고, 사용자2는 음악가 M1을 선호할 때 사용자1과 사용자2의 성향이 같거나 비슷하다면 사용자2에게 M2를 추천하게 된다.

Shardanand(1995)는 피어슨 상관계수를 이용하여 현 사용자와 기존 사용자들 사이의 유사도를 계산하여 사용자의 이웃(neighborhood)을 결정하여 식 (2)를 이용하여 선호도를 예측하였다.

(3) Fab(Balabanovi and Shoham, 1997)

Fab은 스탠포드 대학의 digital library project의 일환으로 개발되었다. Fab 시스템은 협동적 추천과 내용기반 추천을 결합한 웹 기반 추천시스템이다. 추천에 사용되는 사용자의 프로파일은 사용자가 높게 평가한 문서의 단어로 구성되어 있다. 시스템은 특정 주제를 찾아내는 collection agents, 특정 사용자를 찾는 selection agents, 그리고 central router 3개의 컴포넌트(component)로 구성되어 있다. 이 시스템은 내용기반 추천과 협동적 추천을 결합한 시스템으로 다른 사람에게 알려지지 않은 자료를 사용할 수 있다는 내용기반 추천의 장점과 자신이 평가하지 않은 목록에 대한 추천을 가능하게 하는 협동적 추천의 장점을 결합한 것이다. Fab 시스템은 새로운 평가를 반영하기 위해서 매번 시스템을 업데이트해야 한다.

(4) Siteseer(Rucker and Polanco, 1997)

Siteseer는 유사 사용자를 찾거나 사이트를 추천하기 위하여 웹 브라우저의 북마크를 이용하는 추천시스템으로 북마크 목록 중에 중복되는 항목이 있는 사용자를 유사 사용자로 결정하고, 유사 사용자에게 방문하지 않은 사이트를 추천하는 방식이다. 그러나 웹 페이지의 북마크만으로는 목록에 대한 선호도 차이를 알기 어려우므로 이를 바탕으로 하는 추천에는 한계가 있다.

2.2 협동적 추천시스템의 일반적인 구성

추천시스템은 거래되는 상품과 사용자에 따라 구매 데이터가 수집되며 특정 사용자에게 추천을 하는 과정은 유사도 계산단계와 사용자의 선호도를 계산하는 두 단계로 이루어진다 (Billsus *et al*., 1998; Herlocker *et al*., 1999).

유사도 계산단계에서는 사용자의 모든 구매데이터를 기본으로 사용자 간 유사도를 계산한다. 유사도의 계산과정은 <Figure 1>과 같다.

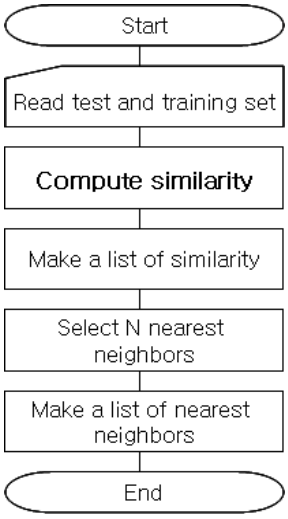

Figure 1. Flowchart of making a list of nearest neighbors.

본 연구에서는 사용자 간의 유사도를 계산하기 위하여 벡터 유사도(vector similarity)를 사용했다. 피어슨 상관계수는 여러 분야에서 많이 사용되고 있지만 사용자의 평가수가 적은 경우 사용자의 평가가 모두 일치하거나 하나의 평가값만 존재할 경우 다른 사용자의 평가에 관계없이 유사도는 0이 된다. 이러한 문제를 피하고자 Breese *et al*.(1998)이 제안한 식 (3)의 벡터 유사도를 사용했다.

$$Sim(Q,D) = \frac{Q \cdot D}{|Q| \cdot |D|} = \frac{\sum_{i=0}^{n} x_{iq} \times x_{id}}{\sqrt{\sum_{i=0}^{n} x_{iq}^2} \times \sqrt{\sum_{i=0}^{n} x_{id}^2}} \quad (3)$$

여기서 Q와 D는 사용자 A와 B가 상품에 대하여 평가한 데이터 벡터를 말한다. 사용자 A가 평가한 데이터 벡터가 $Q = (x_{1q}, x_{2q}, x_{3q}, \ldots, x_{nq})$, 사용자 B가 평가한 데이터 벡터가 $D = (x_{1d}, x_{2d}, x_{3d}, \ldots, x_{nd})$이며, 여기서 x_{1q}는

사용자 A가 상품 x_1에 대하여 평가한 값이고, x_{1d}는 사용자 B가 상품 x_1에 대하여 평가한 값이다.

사용자의 선호도를 계산하는 단계에서는 유사도 계산단계에서 작성된 유사도 목록을 기초로 선호도가 높은 상품순서로 추천하게 된다. 선호도는 사용자가 각각의 상품에 대하여 갖고 있는 선호의 정도이다. 상품에 대한 선호도는 <Figure 2>와 같이 구한다.

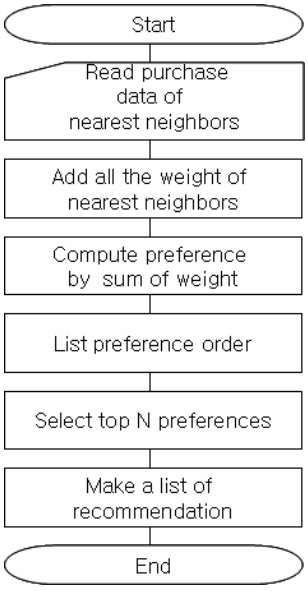

Figure 2. Flowchart of making a list of recommendation by preference.

선호도 예측단계에서 선호도 예측을 위해서 사용될 유사한 사용자의 수를 결정하는 방법으로 사용자 간의 유사도가 일정 값 이상인 사용자들을 이용하는 방법과 특정 사용자와 유사한 n명의 이웃(nearest neighbors)을 이용하는 방법이 있다. 본 연구에서는 특정 사용자와 유사한 n명의 이웃을 이용하는 식 (4)의 예측방법을 이용하였다 (Billsus *et al*., 1998; Breese *et al*., 1998).

$$P_{a,i} = \overline{r_a} + \frac{\sum_{r_{u,i}}^{n} (r_{u,i} - \overline{r_u})\, w_{a,u}}{\sum_{u=1}^{n} w_{a,u}} \qquad (4)$$

여기서, $w_{a,u}$는 사용자 a의 현재 사용자에 대한 유사도 값, $\overline{r_a}$는 사용자 a의 선호도 값의 평균, $r_{u,i}$는 사용자 u의 i번째 상품에 대한 평가값이고, $P_{a,i}$는 사용자 a의 i번째 상품에 대한 예측값이다.

3. 속성추출방법을 적용한 추천시스템 제안

속성추출방법을 적용하여 제안된 추천시스템에서 속성추출 적용단계는 추천시스템에서 유사도 계산단계 이전에 적용되고, 추천에 사용되는 속성 중에서 중요도가 높은 속성을 선택하여 추천시스템에 사용한다. 중요도가 높은 속성만을 사용하게 되므로 유사도의 계산량을 줄임으로써 추천의 속도가 빨라진다. 알고리듬의 계산량은 '시간 복잡도'로서 평가를 하는데 수행시간은 O($n\log n$), O(n), O(n^2) 등으로 나타낸다(Weiss, 1999).

여기서 수행속도 T(n)=O(n^2)은 데이터의 개수가 n일 때 그 수행시간은 n^2에 비례한다. 알고리듬에 사용되는 자료를 줄이는 것이 계산속도에 큰 영향을 주게 된다.

3.1 속성추출(feature selection)방법

기계학습 연구에서 다양한 속성추출방법이 연구되고 있으며 본 연구에서는 문서 빈도수(document frequency)와 엔트로피(entropy)를 이용하였다.

(1) 문서 빈도수

문서 빈도수(document frequency)(Yang and Pedersen, 1997)는 한 개의 속성이 학습자료(training set)에 포함된 빈도로 학습자료에서 각각의 구별되는 속성에 대해 학습자료에서의 빈도수를 계산한 후 정해 놓은 임계값보다 작은 값을 가지는 속성을 제거하여 구한다. 문서 빈도수는 학습자료에 드물게 나타나는 속성이 학습자료의 분류를 예측하는 데 정보력을 가지고 있지 못하다는 기본적인 가정에 기반을 둔다. 임계값을 기준으로 정보력이 낮은 속성을 제거함으로써 속성공간의 차수를 줄일 수 있다. 문서 빈도수 방법은 속성수를 줄일 수 있는 간단한 방법이지만 시행착오적인 방법으로 정보력 있는 속성을 선택하게 된다(Weiss, 1999).

학습자료에서 n명의 사용자 중에서 i번째 사용자 i 에 속성 t가 존재하면 $\text{사용자}_i(t)=1$, 속성 t가 존재하지 않으면 $\text{사용자}_i(t)=0$이 되어 문서 빈도수(DF)는 식 (5)와 같다.

$$DF(t) = \sum_{i=1}^{n} \text{사용자}_i(t) \qquad (5)$$

(2) 엔트로피

엔트로피(Entropy)는 임의의 학습자료의 불순도를 측정하는 수단이다. 엔트로피가 클수록 불순도(impurity)가 높아지고 엔트로피가 0에 가까울수록 순도(purity)가 높아진다.

속성 값이 c개일 경우 엔트로피는 식 (6)과 같다.

$$H(S) \equiv \sum_{i=1}^{c} -p_i \log_2 p_i \qquad (6)$$

여기서 S는 학습자료를 말하고, p_i는 S에 속하는 클래스 i의 비율을 말한다. 타깃 클래스가 c개의 속성값을 갖는 경우 최대 엔트로피는 $\log_2 C$가 된다.

3.2 제안된 협동적 추천시스템

본 연구에서 제안한 속성추출 과정은 다음의 <Figure 3>과 같다.

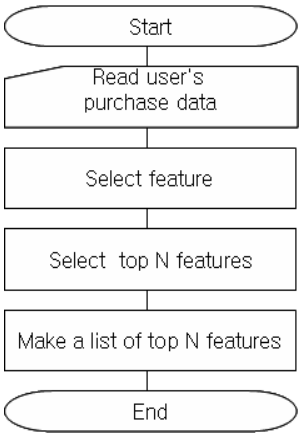

Figure 3. Flowchart of feature selection.

모든 사용자의 구매정보 리스트로부터 구매정보를 읽고 속성추출방법인 문서 빈도수와 엔트로피를 적용하여 중요한 속성을 선택한다. 여러 번의 실험을 통해서 적절한 N개의 속성을 선택하여 중요속성 리스트를 작성하고 속성추출단계를 종료한다. 속성추출방법을 적용함으로써 쇼핑몰에서 판매되는 제품의 전체 구매정보 중에서 유사도 계산에 많은 영향을 미치는 속성만을 선택하여 사용하게 된다. 예를 들면, 구매자가 쇼핑몰에 진열된 100가지의 상품 중에서 많이 판매되는 20가지의 상품을 구입하고, 다른 구매자들도 유사한 상품을 구입했다고 한다면 구매자 간의 유사도를 계산하기 위해서는 많이 판매된 상품의 거래를 비교함으로써 쉽게 찾을 수 있을 것이다.

속성추출방법인 문서 빈도수는 각 상품에 대해 얼마나 많은 사람이 구매를 했는지를 평가하여 많은 사람이 구매한 상품을 중요하다고 평가하며 엔트로피는 불순도를 기준으로 추천시스템에 사용되는 중요속성을 선택한다.

다음의 <Figure 4>는 속성추출단계를 포함한 본 논문에서 제안한 추천시스템의 흐름도를 보여주고 있다.

속성추출 적용 추천시스템은 속성추출단계, 유사도 계산단계, 상품의 선호도 예측단계로 나눌 수 있다. 속성추출단계에서는 앞에서 언급한 속성추출방법인 문서 빈도수와 엔트로피를 이용해서 중요속성을 선택하여 리스트로 작성한다. 유사도 계산단계에서는 속성추출단계에서 얻은 중요속성만을 가지고 각각의 사용자 간 구매성향이 얼마나 유사한가를 비교하기 위해서 벡터 유사도를 이용해서 유사도를 구하고 유사도 리스트를 작성하게 된다. 마지막 단계의 상품의 선호도 예측단계에서는 기준치 이상의 유사 사용자의 자료를 이용해서 각 사용자가 상품에 대해 어떠한 선호도를 가지는지 계산되며 특정 구매자에게 선호도가 높은 상품이 추천된다.

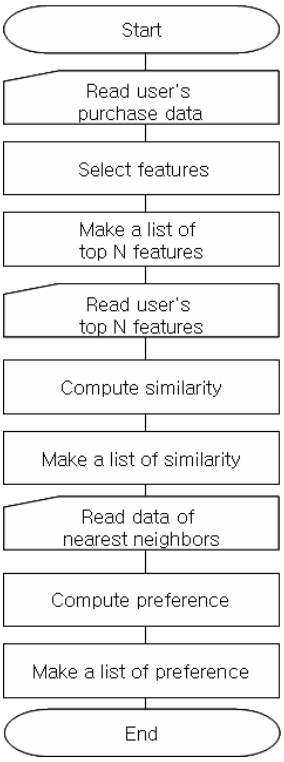

Figure 4. Flowchart of a collaborative recommendation system using feature selection.

4. 제안된 추천시스템의 성능평가

4.1 실험자료

본 실험에서는 미네소타 대학의 GroupLens Research Project에 의해서 1997년 9월부터 1998년 4월까지 약 7개월 동안 수집된 MovieLens 자료(http://movielens.umn.edu)를 사용하였다.

MovieLens 자료는 참여자 943명이 조사대상인 1,682편의 영화중 최소한 20편 이상의 영화에 대하여 1부터 5까지 평가한 약 100,000만 건의 자료로 구성되어 있다.

성능평가실험을 위해 943명 전체의 자료에서 사용자 200명을 무작위로 선택하였으며 선택한 자료의 80%는 추천시스템의 유사도와 제품의 선호도를 구하기 위해서 학습자료로 사용하였고, 자료의 20%는 테스트 자료로 사용하였다. 영화는 1~5의 숫자로 평가가 되었는데 본 실험에서는 4와 5로 평가한 자료는 1로, 1부터 3까지로 평가한 자료는 0으로 바꾸었다. 이는 상품에 대한 선호도를 단순히 구매여부로 나타내기 위한 것이다. 변환된 자료에서 영화를 상품으로 본다면 1은 상품의 구입을 의미하고 0은 구매하지 않은 것을 의미한다.

4.2 실험방법

첫 번째 실험에서 유사 사용자(nearest neighbor)를 몇 명으로

정해야 성능이 가장 좋은지를 알기 위하여 유사 사용자의 수를 단계별로 증가시키면서 기존의 협동적 추천시스템의 성능을 평가했다. 두 번째 실험에서는 속성추출을 적용한 추천시스템에서 중요속성의 수를 단계적으로 증가시키면서 성능을 측정하여 중요한 속성의 수를 파악하였다. 세 번째 실험에서는 기존의 추천시스템과 중요속성을 300개 사용한 속성추출 적용 추천시스템의 연산횟수를 비교하였다.

4.3 성능평가방법

본 실험에서는 추천시스템의 성능을 평가하기 위하여 재현율(recall)과 정확률(precision)을 이용하였다(Balabanovi and Shoham, 1997). 추천개수를 선호도가 높은 10개를 선택하는 Top-10 추천을 했고, 10개의 추천목록에 대해서 재현율과 정확률을 평가했다.

재현율은 추천 리스트와 테스트 자료의 중복된 개수(number of hits)를 테스트 자료의 개수로 나눈 값이다.

$$\text{재현율(recall)} = \frac{\text{추천리스트와 테스트자료의 중복된 개수}}{\text{테스트자료의 개수}} \quad (7)$$

정확률은 추천의 정확성을 평가하기 척도로 추천 목록의 정확성이 어느 정도 정확한가를 평가할 수 있다. 정확률은 추천리스트와 테스트 자료의 중복된 개수(number of hits)를 추천 리스트의 개수로 나눈 값이다.

$$\text{정확률(precision)} = \frac{\text{추천리스트와 테스트자료의 중복된 개수}}{\text{추천리스트의 개수}} \quad (8)$$

4.4 실험결과

200명이 1682편의 영화에 대하여 평가한 자료를 이용해서 기존의 협동적 추천시스템과 개선된 추천시스템의 성능비교 실험을 하였다.

<Table 1>은 속성추출방법을 적용하지 않은 기존의 추천시스템의 재현율이다. 사용자 간의 유사도를 구하고, 유사도가 비슷한 사용자 상위 10명, 20명, 30명 등, 10명씩 단계별로 증가하도록 하여 모든 유사 사용자가 포함될 때까지 증가시키면서 재현율을 구했다.

Table 1. Recall of a collaborative recommendation system

	NN 10	NN 20	NN 30	NN 40	NN 50	NN 60	NN 80	NN 100	NN 120	NN 160	NN 200
recall	0.177	0.185	0.177	0.177	0.187	0.174	0.174	0.160	0.153	0.145	0.145

(NN : Number of nearest neighbors)

<Table 1>의 내용을 그래프로 나타내면 <Figure 5>와 같으며 단계별로 증가했을 경우 재현율의 변화를 알 수 있다.

<Figure 5>에서 유사도가 비슷한 유사 사용자를 50명 선택한 경우 재현율이 가장 좋다는 것을 알 수 있다. 그리고 유사 사용자를 100명 이상 선택했을 경우 재현율이 계속 감소하는 것을 알 수 있다. <Figure 5>에서 적정 수준의 유사 사용자를 사용할 경우가 모든 유사 사용자를 사용한 경우보다 재현율이 좋다는 것을 알 수 있다.

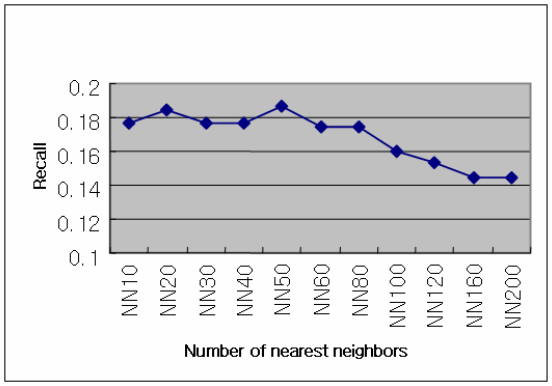

Figure 5. Recall of a collaborative recommendation system.

<Table 2>와 <Figure 6>은 협동적 추천시스템의 유사도 계산단계에서 속성추출기법인 문서 빈도수(document frequency)를 적용했을 경우의 실험결과이다. 실험에서 문서 빈도수를 이용해서 전체 속성 1,682개 중 중요속성을 단계적으로 증가시키면서 재현율을 구했다. 추천시스템의 재현율이 가장 좋았던 유사 사용자 수가 50명일 때 300개의 중요속성을 선택한 경우에 재현율이 0.193으로 가장 우수했다.

Table 2. Recall when using feature selection(document frequency)

Number of feature	100	300	500	700	1000	total
NN50	0.174	0.193	0.184	0.179	0.180	0.167

또한 모든 속성을 사용하는 기존의 추천시스템보다 속성추출을 통해서 중요속성을 선택하여 추천시스템에 적용하는 것이 더 우수한 성능을 보여주고 있다.

엔트로피를 적용하여 협동적 추천시스템의 성능을 평가한

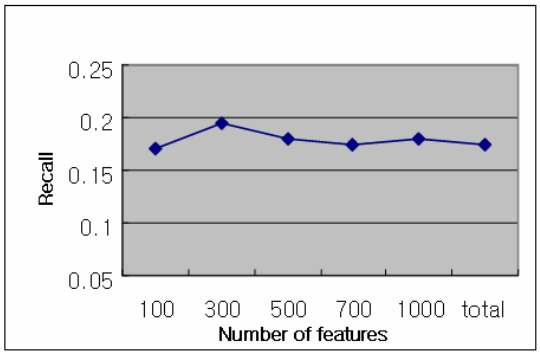

Figure 6. Recall when using feature selection(document frequency).

실험도 문서 빈도수를 적용한 실험과 동일하게 하였고, 중요속성의 선택을 위해서 엔트로피를 이용했다. 실험의 결과는 <Table 3>, <Figure 7>과 같다.

Table 3. Recall when using feature selection(entropy)

Number of Feature	100	300	500	700	1000	total
NN50	0.174	0.193	0.184	0.179	0.180	0.167

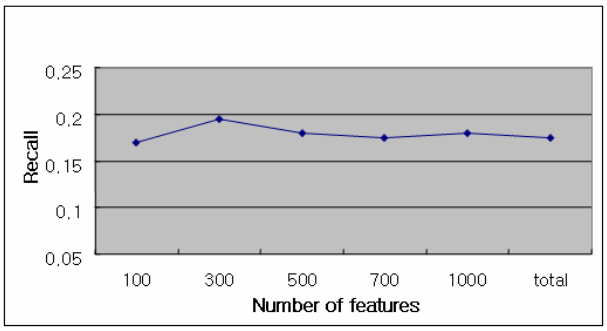

Figure 7. Recall when using feature selection(entropy).

속성추출방식으로 문서 빈도수와 엔트로피를 이용한 경우 실험결과가 동일하게 나온 이유는, 두 가지 방법 모두 거의 동일한 속성들이 선택되었기 때문이다. 이러한 결과는 자료의 양이 적고 영화의 관람 여부를 측정한 자료이므로 인기 있는 영화가 빈도가 높게 나오고 마니아층을 이루는 사람의 자료에는 비슷한 유형의 자료가 많이 포함되어 있는 실험자료의 특성에 기인한다고 볼 수 있다.

<Figure 8>은 유사 사용자를 50명으로 제한하고 문서 빈도수를 이용해서 중요속성 수를 단계적으로 증가시키면서 정확률을 측정한 결과다. 실험에서는 중요변수가 300개인 경우에 정확률이 가장 좋은 것을 알 수 있다. 기존의 협동적 추천시스템은 전체 속성을 모두 이용하기 때문에 전체의 속성을 이용했을 때의 정확률과 같다.

실험결과에서와 같이 속성추출을 하지 않은 경우보다 속성

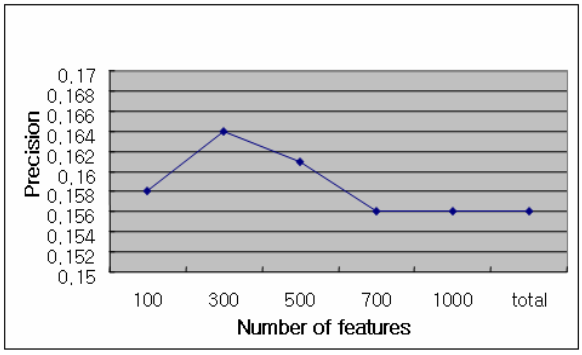

Figure 8. Precision when using feature selection (document frequency).

추출을 적용한 추천시스템의 성능이 많이 향상됨을 알 수 있다. 속성추출을 적용한 추천시스템의 속도가 얼마나 빨라졌는지 알기 위해 기존의 방법과 유사도계산의 연산횟수를 비교하였다.

<Figure 9>는 기존의 추천시스템과 속성추출방법을 적용한 경우 추천시스템 중에서 유사 사용자의 수와 관계없이 재현율이 우수했던 문서 빈도수를 적용해서 300개의 중요속성을 택하여 200명의 유사도를 계산한 경우의 연산횟수를 비교하여 보여주고 있다.

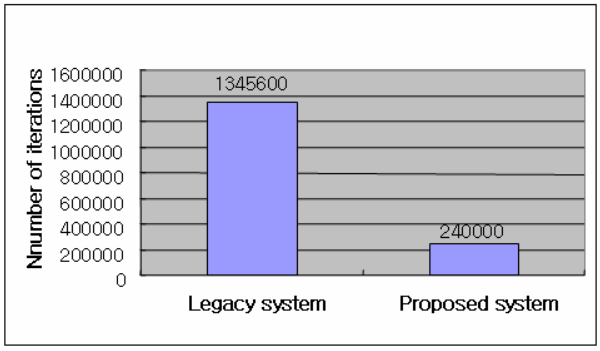

Figure 9. Comparison of the number of iterations.

<Figure 9>에서와 같이 속성추출도 속성이 개수가 줄어든 경우의 연산횟수가 현저히 줄어든 것을 알 수 있다. 본 논문에서 제안한 추천시스템의 유사도 계산단계의 수행속도는 T(n)= O(n^2) 이다(Weiss, 1999). 즉 유사도계산에 사용되는 속성의 수가 5배 감소했으므로 계산속도는 25배 정도가 감소하게 된다. 계산속도가 빨라짐으로써 사용자와 상품수가 늘어남에 따라 추천의 속도가 느려지는 확장성의 문제를 해결할 수 있음을 알 수 있다.

5. 결 론

전자상거래의 급성장으로 효율적이고 효과적인 추천시스템 개발의 중요성이 날로 증대되고 있다. 본 연구는 협동적 추천시스템의 성능을 향상시키고, 확장성 문제를 해결하기 위하여 속성추출방법을 적용한 추천시스템을 제안하였다.

제안된 추천시스템에서는 추천을 위한 유사도를 계산단계에서 연산량을 줄임으로써 추천시스템 성능을 증가시키면서도 추천시간을 단축할 수 있었다. 추천시간의 단축으로 사이트에서 사용자와 상품의 수가 증가하여 추천의 속도가 느려지는 확장성 문제를 해결할 수 있게 되었다. 본 실험에서 사용된 자료의 특성상 단어 빈도수와 엔트로피를 이용한 속성추출방법으로 동일한 결과를 얻었지만 다양한 속성추출방식으로 다른 형태의 자료를 이용할 경우 다른 의미 있는 결과를 얻을 수 있을 것이다.

본 연구가 실용화되기 위해서는 다양하고 보다 많은 자료를 이용한 반복실험으로 시스템의 안정성과 범용성을 확보해야 할 것 이다.

참고문헌

Kim, S. (2001), Dimensionality reduction to solve the data sparseness problem in recommendation systems, Master's Thesis, Dongguk Univ.

Jung, E.(2002), Scheme for accuracy enhancement of collaborative recommendation system, Master's Thesis, Hongik Univ.

Jo, S. (2002), User simility measurement method using entropy and defult voting in recommendation system, Master's Thesis, Inha Univ.

Balabanovi,M., and Shoham, Y.(1997), Content-based Collaborative Recommendation , Communications of the ACM , 40.

Basu, C., Hirsh, H. and Cohen, W. (1998), Recommender Systems. Recommendation As Classification: Using Social And Content-Based Information, Proceedings of the Workshop on Recommendation Systems. AAAI Press.

Billsus, D. and Pazzani, M. J. (1998), Learning Collaborative Information Filters, In Proceedings of Recommender Systems Workshop. Tech. Report WS-98-08, AAAI Press.

Breese, John S., Heckerman, David and Kadie, Carl(1998), Empirical Analysis of Predictive for collaborative Filtering, Proceedings of the 14th Conference of Uncertainty in Artificial Intelligence.

Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D. and Sarti, M.(1999) Combining Content-Based and Collaborative Filters in an Online Newspaper, ACM SIGIR Workshop on Recommender Systems, Berkeley, CA.

Herlocker, J. L., Konstan, J. A., Borchers, A. and Riedl, J. (1999), An Algori-thmic Framework for Performing Collaborative Filtering, Proceedings of the Conference on Research and Development in Information Retrieval.

Karypis, G.(2000), Evaluation of Item-Based Top-N Recommendation Algorithms, Technology Report CS-TR-00-46, Computer Science Dept., University of Minnesota.

Konstan, J. A., Miller, B. N., D. Maltz, Jerlocker, J. L., Gordon, L. R. and Riedl, J. (1997), GroupLens: Applying Collaborative Filtering to Usenet News, Communications of the ACM.

Mitchell, T. M.(1997), MACHINE LEARNING, The McGraw-Hill Company.

Resnick, P., Iacovou, N., Suchak, M., Pergstom, P. and Riedl, J.(1994), GroupLens: An Open Architecture for Collaborative Filtering of Netnews, Proceedings of the ACM 1994 Conference on Computer Supported Cooperative Work.

Rucker, J. and Polanco, M. J.(1997), Sitesser: Personalized navigation on the Web, Communications of the ACM.

Sarwar, B., Karypis, G., Konstan, J. A. and J.Riedl(2000), Application of Dimensionality Reduction in Recommender System-A Case Study, In WebKDD 00-Web-mining for E-Commerce Workshop.

Sarwar, B., Karypis, G., J. Konstan, and Riedl, J. (2001), Irem-based Collaborative Filtering Recommendation Algorithms, Accepted for publication at the WWW10 Conference.

Schafer, J. Ben, Konstant, J. and Reidl, J. (1999), Recommender Systems in E-Commerce, In Proceedings of ACM E-Commerce 1999 conference.

Shardanand, U.(1995), Social information filtering for music recommendation, Technical Report MA95, MIT Media Laboratory.

Weiss, Mark Allen(1999), Data Structures & Algorithm Analysis in Java, Addison-Wesley Pub Co.

Yang, Y. and Pedersen, J.(1997), A comparative study on feature selection in text categorization, ICML.

National Statistical Office, http://www.nso.go.kr:7001/main.cfm

MovieLens Dataset, http://www.grouplens.org/data/

Ministry of Information and Communication (2005), IT Report 2005.

Korea Institute For Electronic Commerce (2005), e-Business white book.

유 상 종

동국대학교 산업공학 학사

동국대학교 산업공학 석사

현재: (주)소프트온모바일 재직

관심분야: 전자상거래, 데이터마이닝

권 영 식

서울대학교 산업공학 학사

한국과학기술원 산업공학 석사

한국과학기술원 산업공학 박사

현재: 동국대학교 산업시스템공학과 교수

관심분야: 데이터마이닝, 산업공학