SVD을 이용한 저차원 공간에서 협력적 여과

정 준† · 이 필 규††

요 약

추천 시스템은 구매할 상품을 사용자가 찾는 것을 도와 주는 시스템이다. 추천 시스템에서 사용되고 있는 여러 가지 방법 중에 대표적인 방법인 협력적 여과는 유사한 사용자들에 기초하여 그 사용자들이 선호하는 상품을 교차 추천해주는 방법이다. 사용자들에 대한 정보는 상품을 평가한 등급에 기초하고, 유사한 사용자는 평가 패턴의 유사성으로 판단된다. 순수한 협력적 여과는 사용자가 증가함에 따라서 평가 자료의 차원이 증가한다. 평가 자료의 고차원성은 자료의 희소성을 증가시켜 협력적 여과의 성능이 저하되는 문제점을 가지고 있다. 따라서, 본 논문에서는 SVD를 이용하여 평가 자료의 차원을 감소시켜 희소성을 최소화하는 방법을 고찰하며, 협력적 여과에 미치는 영향을 실험적으로 제시한다. 결과적으로 SVD를 이용한 협력적 방법은 순수한 협력적 여과 방법과 비교하여 충분히 정확한 성능을 보였다.

A Collaborative Filtering using SVD on Low-Dimensional Space

Jun Jeong† · Pil-Kyu Rhee††

ABSTRACT

Recommender System can help users to find products to purchase. A representative method for recommender systems is collaborative filtering (CF). It predict products that user may like based on a group of similar users. User information is based on user's ratings for products and similarities of users are measured by ratings. As user is increasing tremendously, the performance of the pure collaborative filtering is lowed because of high dimensionality and scarcity of data. We consider the effect of dimension deduction in collaborative filtering to cope with scarcity of data experimentally. We suggest that SVD improves the performance of collaborative filtering in comparison with pure collaborative filtering.

키워드 : 협력적 여과(Collaborative Filtering), 추천 시스템(Recommendation System), SVD(Singular Value Decomposition), 차원 감소(Dimension Deduction)

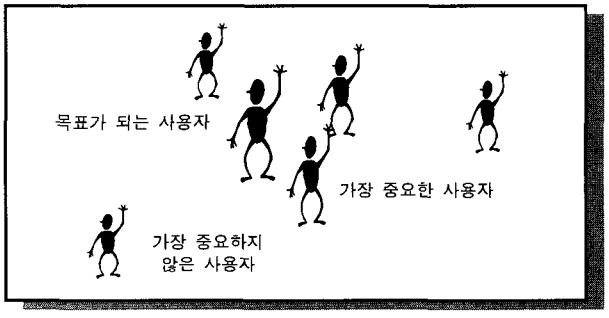
1. 서 론

정보 산업은 인터넷의 폭발적인 확산과 더불어 이제 는 정보의 혁명이라고 할 수 있을 정도로 급속하게 발전해 왔다. 일상 생활에서 접하는 정보뿐만 아니라 인터넷을 통한 정보 획득도 중요한 수단이 되고 있으며 점차 그 중요성은 더욱 강조되고 있다[1, 9].

또한 인터넷을 이용한 전자 상거래도 보편화도 되어 왔다. 전자 상거래의 발전과 더불어 추천 시스템(recommender system)의 도입이 활성화되어 왔으며, 대표적으로 아마존(www.amazon.com)과 같은 사이트를 예로 들 수 있다. 추천 시스템이란 사용자로부터 정보를 수집하여 그 사용자가 선호할 만한 상품을 예측하는 시스템이며, 내용 기반 여과(content-based filtering), 협력적 여과(collaborative filtering), 에이전트(agent), 규칙 기반 여과(rule-based filtering) 등과 같은 기술들이 추천 시스템에서 주요하게 사용되어 왔다[9, 10].

※ 본 연구는 2001년 인하대학교 교내 연구비의 지원으로 연구되었음.
† 준 회 원 : 인하대학교 대학원 전자계산공학과
†† 종신회원 : 인하대학교 전자계산공학과 교수
논문접수 : 2002년 10월 23일, 심사완료 : 2003년 4월 22일

추천 시스템에서 대표적으로 사용되는 협력적 여과 방법은 "구전"이라는 사회 현상을 자동화한 방법이다. 일상적으로 영화, 책, 음악 등과 같은 것을 선택할 때, 우리는 다른 사람들의 선호도 정보를 이용하여 판단한다. 이러한 자연적인 사회 현상 과정을 도와 주고, 보강해주는 것이 협력적 여과이다.



(그림 1) 협력적 여과

협력적 여과는 상품에 대한 선호도를 예측하기 위하여, 사용자로부터 일정 개수이상 상품에 대해서 평가 자료(rating)를 입력받는다. 평가는 일반적으로 5~7단계를 사용한다.

예를 들어, 사용자가 각 상품에 대하여 자신이 선호하는 정도를 1에서 7까지 7단계로 평가한다고 하자. 여기서 1을 가장 낮은 선호도로 7을 가장 높은 선호도로 할 때, 평가 정보는 <표 1>과 같이 나타낼 수 있다. 사용자 2의 평가 정보를 보면 상품 1에 대해서는 좋게 평가했고 상품 4에 나쁘게 평가한 것을 알 수 있다. 따라서, 이러한 평가 정보는 사용자들간의 유사성을 측정할 수 있는 자료가 된다. 평가 자료를 기반으로 사용자의 유사성은 상관계수(correlation), 벡터의 유사성(vector similarity), 신경망(neural network), 베이지안 네트워크(baysian network) 등과 같은 방법을 이용하여 측정하며, 유사성을 이용하여 개인화된 유사한 사용자 그룹을 생성한다. 마지막으로 생성된 그룹 안에서 사용자가 평가하지 않은 상품에 대해서 상호 교차 추천해준다.

〈표 1〉 사용자 평가 정보

	상품 1	상품 2	상품 3	상품 4
사용자 1	7	4		
사용자 2	7	5		1
사용자 3		2	3	

이러한 협력적 여과 방법은 사용자의 수가 증가함에 평가 자료의 희소성이 증가하기 때문에 성능이 감소되는 성향을 가진다. 따라서, 고차원 평가 자료를 효과적으로 처리하는 방법이 필요하며, 본 논문에서 Singular Value Decomposition(SVD)을 이용한 차원 감소 방법을 고찰한다. SVD를 이용한 차원 감소 방법에서 중요한 점은 감소될 차원을 결정하는 것이다. 실험적으로 감소될 차원을 결정하기 위하여 협력적 여과를 위한 대표적인 실험자료인 Eachmovie 자료와 Movielens 자료를 이용하여 실험을 수행하였다. 또한 전반적으로 성능이 우수하다고 알려진 상관계수법을 이용한 협력적 여과 방법과 비교평가를 수행하였다.

본 논문은 다음과 같이 구성된다. 2절에서는 협력적 여과와 관련된 연구를 소개하며, 3절에서는 본 논문에서 고찰하는 방법에 대한 기본적인 소개와 전체적인 방법을 설명하며, 4절에는 실험 방법 및 결과에 대해서 제시하며, 5절에는 결론을 제시한다.

2. 관련 연구

Tapestry[3]는 최초의 협력적 여과 시스템(collaborative filtering system)으로서 정보 과잉의 문제를 해결하기 위하여 작은 그룹의 사람들이 함께 작업하는 것을 도와주는 시스템이다. 메일, 뉴스그룹에서 적용되었으며 텍스트 문서 형식의 주석과 수치적인 평가 값과 불리언 평가 값으로 입력 정보에 대한 선호도를 주었으며, 내용기반 여과 방법 및 주관적인 평가 방법을 통해서 소규모 사용자들에게 링크, 아이템, 평가 값 등을 제공하여 준다. 사용자가 추천을 받기 위해서는 TQL이라는 SQL과 유사한 언어로서 자신이 필요로 하는 것을 기술해야 한다. 그러나, Tapestry는 몇 가지 단점을 가지고 있다. ① 가장 가까운 이웃을 찾아 주는 기능이 없기 때문에 누가 나와 비슷한 취향을 가지고 있는지 먼저 알아야 한다. ② 효과적인 결과를 받기 위하여 능동적으로 주석을 달아야 한다. ③ 소규모 사용자들을 대상으로 한다. ④ 필요한 정보를 표현하는 언어가 사용하기가 어렵다. ⑤ 우연한 발견을 지원하지 않는다.

GroupLens[8]는 기사 중에서 다른 사용자들의 관심을 예측하기 위하여 사용자들로부터 단계적 평가를 수집하고 보급하고 사용하기 위한 분산 시스템이다. 이 시스템은 단계적 평가를 수집하고 예측하는 "Better Bit Bureaus"와 뿐만 아니라 뉴스를 읽을 수 있는 클라이언트를 포함한다. GroupLens 설계의 목표는 기존의 뉴스를 읽는 클라이언트와 쉬운 통합 방법과 평가를 하기 위한 편리한 방법을 제공하는 것과 사용자의 기사에 대한 평가 값의 예측을 제공하는 것이다. 사용자의 취향 분석을 위하여 각 사용자가 뉴스에 대해서 5단계로 선호도를 평가를 하거나 시스템이 사용자가 뉴스를 읽는 시간을 모니터링을 한다. 사용자간의 유사성을 측정하기 위하여 Ringo과 비슷하게 Pearson r 방법을 사용하였다.

Ringo[5]는 웹과 메일을 통해서 음악가에 대한 명시적 평가(explicit rating)로 구성된 사용자 프로파일에 기초한 음악 추천 시스템이다. 사회적인 정보 여과(social information filtering)이라는 용어를 처음으로 사용하였다. 음악가에 대한 평가는 5단계로 이루어지며 유사한 사용자들을 찾기 위하여 통계학에서 상관계수(correlation coefficient)를 찾기 위한 방법인 Pearson r을 수정하여 사용하였다. 계산된 상관계수에 가중치를 주어서 임계값(threshold) 이상의 유사성을 가지는 사용자들을 이웃이라고 판단하여, 그 이웃들을 기초하여 회귀분석을 통해서 음악가에 대한 선호도를 측정하였다.

3. SVD를 이용한 협력적 여과

3.1 Singular Value Decomposition(SVD)

SVD[12]는 singular하거나 수치적으로 singular한 행렬이나 방정식의 집합을 처리하기 위한 매우 강력한 방법이다. Gaussian elimination과 LU 분해가 만족할 만한 결과를 주지 못할 때 SVD는 문제가 무엇인지 정확하게 분석할 수 있다. 어떤 경우에는 SVD는 문제를 분석할 수 없지만, 문제를 풀 수는 있다. 또한 SVD는 대부분의 linear least-squaure문제를 풀기 위한 방법이다.

다음은 SVD의 형식적인 정의를 설명한다.

$A = U\Sigma V^{-1}$과 같은 분해는 어떤 크기의 행렬에서도 가능하며, 이러한 형식의 분해 중에서, SVD는 응용 선행 대수에서 가장 유용한 행렬 분해이다.

SVD는 직각 행렬을 위해서 모방될 수 있는 다음과 같은

상미분 방정식의 대각화의 속성에 기초한다. 대칭 행렬 A의 고유값(eigenvalue)의 절대값은 A가 어떤 벡터(eigenvector)를 연장하거나 축소를 시킬 수 있는 양을 측정한다. 만약 $Ax = \lambda x$이고 $\|x\| = 1$이라면,

$$\|Ax\| = \|\lambda x\| = |\lambda| \|x\| = |\lambda| \quad (1)$$

만약 λ_1이 가장 중요한 고유값이라면, 상응하는 고유벡터 v_1은 A의 연장 결과가 가장 크게 작용할 수 있는 방향을 결정한다. 즉, Ax의 길이는 식 (1)에 의해서 $x = v_1$이고 $\|Av_1\| = |\lambda|$일 때 최대화된다. v_1과 $|\lambda_1|$의 기술은 singular value decomposition에 이르는 직각 행렬에 대한 유사물을 가진다.

A를 m×n행렬이라고 하면, A^tA는 대칭이고 직각으로 대각화 된다. $\{v_1, \cdots, v_n\}$을 A^tA의 고유값을 구성하는 n차원 공간(R^n)상의 직각 기초라고 하고, $\lambda_1, \cdots, \lambda_n$을 A^tA의 관련된 고유값이라고 하면, $1 \le i \le n$상에서,

$$\begin{aligned} \|Av_i\|^2 &= (Av_i)^T Av_i = v_i^T A^T A v_i \\ &= v_i^T(\lambda_i v_i) \quad v_i\text{는 } A^TA\text{의 고유벡터} \\ &= \lambda_i \quad v_i\text{는 단위 벡터} \end{aligned} \quad (2)$$

그래서, A^TA의 고유값은 모두 음수가 아니다. 필요하면 번호를 달아서 고유값을 다음과 같이 정리할 수 있다.

$$\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_n \ge 0 \quad (3)$$

A의 singular value는 $\sigma_1, \cdots, \sigma_n$으로 표시되는 A^TA의 고유값의 제곱근이고, 내림차순으로 정리된다. 즉, 식 (*)에 의해서 $1 \le i \le n$인 경우 $\sigma_1 = \sqrt{\lambda_i}$이다. A의 singular value는 벡터 $Av_1, \cdots, Av_n$의 길이이다.

우선, A라는 행렬이 존재하면 행렬 A의 분해는 식 (4)와 같은 형식의 m×n 대각 행렬 Σ를 포함한다.

$$\Sigma = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} \quad (4)$$

여기서, (m-r)×(n-r) 행렬 D는 m과 n에서 작은 값을 넘지 않는 어떤 계수 r에 대해서 r×r인 대각 행렬이다.

[정리 1] A는 계수가 r인 m×n 행렬이라고 하면, 식 (5)와 같이 m×n 행렬인 Σ가 존재하며, D는 행렬 A의 r개의 singular value이고, $\delta_1 \ge \delta_2 \ge \cdots \ge \delta_r > 0$ 이 되고, 식 (*)과 같은 m×m 직교 행렬인 U와 n×n 직교 행렬 V가 존재한다.

$$A = U\Sigma V^T \quad (5)$$

직교 행렬 U와 V와 Σ를 포함하는 행렬 A의 인수분해인 $U\Sigma V^T$을 행렬 A의 SVD(singular value decomposition)이라고 한다. 행렬 U와 V는 유일하지 않지만, Σ의 대각 원소들은 행렬 A의 singular value이다.

3.2 평가 자료의 차원 감소

u명의 사용자와 i개의 상품을 나타내는 u×i 행렬인 R은 3가지 다른 행렬로 분해될 수 있다.

$$R = U\Sigma V^T \quad (6)$$

여기서, U와 V는 직각 행렬이고, Σ는 대각 행렬이다. 이러한 형태의 분해를 R 행렬의 Singular Value Decomposition(SVD)이라고 한다. U와 V는 각각 좌측, 우측 singular vector라고 하며, Σ는 singular value의 대각 행렬이다.

일반적으로, SVD는 더 작은 행렬을 이용하여 최적의 근사형을 위한 단순한 방법을 제공한다. Σ의 singular value가 크기에 의해서 정렬된다면, k개의 가장 큰 값들을 선택하고 나머지 작은 값들은 0으로 설정한다면, 행렬의 곱은 R에 근사적으로 동일한 행렬인 R′을 생성할 수 있다.

$$R' = U_k \Sigma_k V_k^T \approx R \quad (7)$$

R의 k개의 가장 큰 독립적인 선형 요소를 포함하는 것은 R′가 자료의 중요하게 관련된 구조를 생성하고 대부분의 노이즈를 제거하는 효과를 가진다.

k의 선택에 있어서는 일반적으로 최적의 해는 존재하지 않으나, 실험에 의해서 적절한 수를 결정할 수 있다. 또한 정렬된 singular value에서 k개의 값을 선택하는 방법은 여러 가지가 존재 할 수 있다. 예를 들면, 크기 순서대로 k개를 선택할 수도 있고, 처음 몇 개는 건너뛰고, 그 다음부터 k개를 선택할 수 있다. 이러한 singular value의 수와 선택 방법은 시스템의 성능에 중요한 요소가 된다.

3.3 선호도 예측

상품에 대한 선호도 예측은 평가 자료를 가지는 R 행렬로부터 시작한다. R 행렬은 SVD를 이용하여 분해되여 U, Σ, V^T를 된다.

$$\begin{aligned} R &= U\Sigma V^T \\ RV &= U\Sigma V^T V \\ RV &= U\Sigma(V^T V) \\ RV &= U\Sigma \end{aligned} \quad (8)$$

평가 행렬 R에 V를 곱한 것은 사용자의 특징 벡터에 각 특징들의 중요도를 곱한 효과를 가지게 된다. 즉, 상품을 평가한 값을 직접 비교하여 사용자간의 유사성을 측정하는 것이 아니라, 사용자를 차원이 감소된 저차원 공간으로 사상시켜서 비교하는 효과를 가지게 된다. 따라서, U와 S의 행렬의 곱은 두 사용자의 유사성을 나타내는 벡터를 구성한다.

임의의 두 사용자 i, j의 유사성을 측정하는 방법으로 RV 혹은 $U\Sigma$ 행렬의 i번째 행과 j번째 행의 코사인 거리로 측정되고, 코사인 거리가 클수록 유사성이 크다고 간주된다.

사용자 i에 대한 상품 j에 대한 선호도 예측은 두 가지 방

법이 있다. 첫 번째로, 사용자의 유사성에 대한 경계 값을 설정하여, 그 경계 값 이상의 유사성을 가지는 사용자들을 유사한 사용자로 간주하여 그 사용자들의 평가 값을 가중치가 부여된 평균으로 구할 수 있다. 두 번째로 사용자의 유사성을 기반으로 n명의 상위 사용자들을 유사한 사용자로 간주할 수 있다. 본 논문에서는 첫 번째 방법을 사용한다.

전체적인 과정을 설명하기 전에 필요한 기호를 먼저 설명한다. R =은 모든 n명의 사용자들이 m개의 상품에 대한 평가의 집합이고, R_i =는 사용자 i가 모든 상품 j에 대한 평가 값의 프로파일이다. $\overline{R_i}$은 사용자 i에 대한 평가 값의 평균이다. c_{ij}는 i번째 사용자와 j번째 사용자의 유사성을 나타내고, w_{ij}는 i번째 사용자와 j번째 사용자의 가중치가 부여된 유사성을 나타낸다. T는 임의의 유사성 임계치를 나타낸다.

① 행렬 R을 SVD를 이용하여 분해한다.

$$R = U\Sigma V^T \quad (9)$$

② 임의의 계수 k에 따라 U와 Σ를 이용하여 R′을 구성한다.

$$R' = U\Sigma \quad (10)$$

③ R′을 이용하여 모든 사용자들간의 코사인 거리를 계산한다.

$$c_{ij} = \frac{R_i \cdot R_j}{|R_i||R_j|} \quad (11)$$

④ 사용자들간의 가중치를 계산한다.

$$w_{ij} = \left(\frac{c_{ij} - T}{1-|T|}\right)^2 \quad (12)$$

⑤ 임의의 경계 값 T보다 큰 코사인 거리를 가지는 사용자들을 사용자 i의 유사한 사용자 그룹인 N_i로 간주하고, 사용자 i가 평가하지 않은 상품을 j라 하면, 선호도를 p_{ij}는 가중치가 부여된 평가 값의 평균으로 구한다.

$$p_{ij} = \overline{R_i} + \frac{\sum_k^{N_i}(w_{ik} \times p_{dj})}{\sum_k^{N_i} w_{ik}} \quad (13)$$

(그림 2) 전체적인 과정

4. 실 험

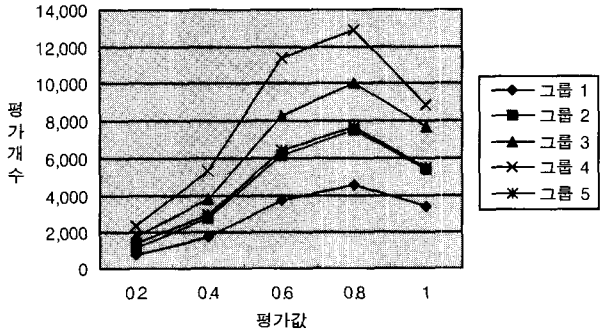
4.1 실험 자료

실험 자료는 DEC Systems Research Center에서 제공하는 EachMovie collaborative filtering data set[2]과 미네소타 대학의 GroupLens Research Project에 의해서 수집된 MovieLens [6]를 사용하였다.

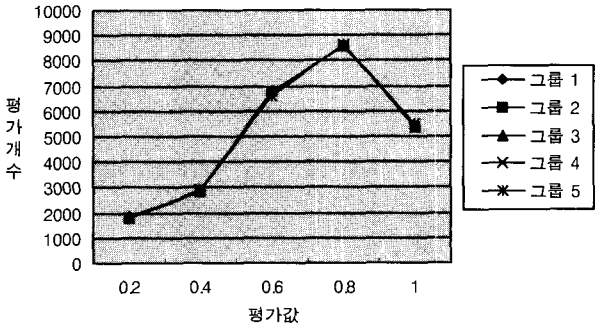
DEC는 18개월 동안 협력적 여과 알고리즘을 실험하기 위하여 EachMovie 추천 서비스를 실행하였다. 그 결과로 수집된 EachMovie 자료는 72,916명의 사용자들이 1628개의 영화와 비디오에 대해서 2,811,983개의 평가 값을 가지고 있고, 사용자의 중요한 정보가 제거되어 협력적 여과 알고리즘에 쉽게 적용될 수 있도록 가공하여 제공되어 있다. 사용자의 평가 정보는 {0.2, 0.4, 0.6, 0.8, 1.0}로 이루어져 있다.

Movielens는 943명의 사용자에 1682개의 영화에 대해서 {1, 2, 3, 4, 5}로 평가한 자료이며, 총 평가 수는 100,000개이다. 각 사용자는 적어도 20개 이상을 평가를 수행했으며, 나이, 성별, 직업, 우편번호와 같은 단순한 인구통계학적인 정보를 포함하고 있다.

효과적인 실험을 위하여 전체 Eachmovie 자료에서는 1000명의 사용자를 임의로 추출하여 사용하였다. Eachmovie 자료에서 사용자들을 추출할 때, 평가수에 따라, 11~20개, 21~30개, 31~40, 41~50개, 11~50를 평가한 5개의 그룹을 추출하였으며, 각 그룹에 대해서 80%는 학습 자료로 나머지 20%는 검사자료를 사용하기 위하여 분리하였다. 또한, Movielens 자료는 전체 자료 중에 80%를 학습 자료로 20%를 검사자료로 분리한 5개의 그룹을 제공하는 것을 이용하였다. (그림 3)과 (그림 4)는 각 실험 자료에 대하여 평가 값에 따라 평가 개수를 나타낸 것이다. (그림 3)에서 보여지듯이 대체로 사용자들은 긍정적인 평가를 가장 많이 내렸으며, 각 그룹의 평가 개수의 차이는 실험 자료를 추출할 때, 각 사용자의 평가 개수를 범위를 한정했기 때문이다. (그림 4)도 (그림 3)과 같이 사용자들은 긍정적인 평가를 가장 많이 내렸으며, 평가 개수가 거의 일치하는 것은 전체 실험 자료에서 같은 비율로 5번 반복하여 추출하였기 때문이다.



(그림 3) 평가값에 따른 평가 개수의 분포(Eachmovie)



(그림 4) 평가값에 따른 평가 개수의 분포(MovieLens)

4.2 평가 방법

알고리즘의 성능을 평가하기 위하여 MAE(Mean Absolute Error)를 사용하였다. MAE은 부호와 관계없이 각각 오차 크기의 평균이며, MSE(Mean-Squared Error)의 값이 작

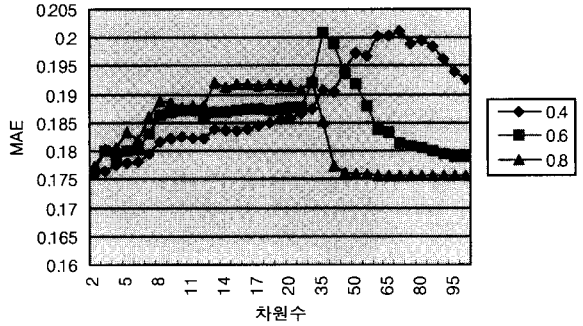
을수록 더 정확한 방법이다. MAE는 다른 오차보다 더 큰 예측 오차의 결과를 강조하는 MSE(Mean-Squared Error)의 특성에 영향을 받지 않으며, 오차의 모든 크기는 그들의 오차의 크기에 따라서 동일하게 취급되어진다.

$$\frac{|p_1 - a_1| + \cdots + |p_n - a_n|}{n} \quad (14)$$

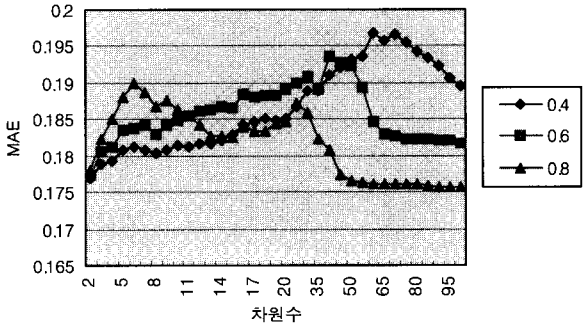
여기서, p_n은 선호도에 대한 예상된 값이고, a_n은 실제 사용자가 평가한 값이 된다. n은 a, p의 총 개수이다.

4.3 실험 방법

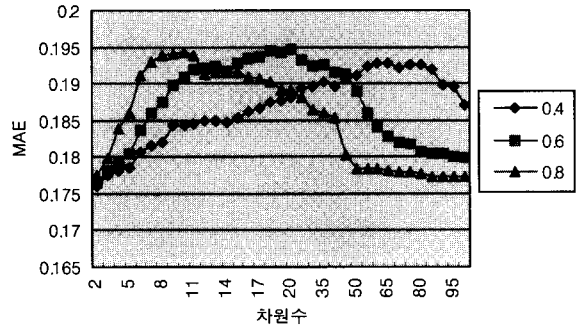
협력적 여과는 사용자의 평가 정보를 이용하여 사용자간의 유사성을 측정하고, 선택적으로 유사성에 가중치를 부여하며, 정해진 임계치 이상의 유사성을 가진 사용자를 유사한 사용자로 구별하여 각 사용자마다 개인화된 유사 사용자 그룹을 생성하여, 생성된 그룹에서 사용자가 평가하지 않은 상품에 대한 선호도를 예측한다. 본 논문은 사용자간의 유사성을 측정하는데 있어서 원래의 평가 정보에서 측정하는 것이 아니라, 평가 정보의 차원수를 감소시켜서 사용한다. 따라서, 감소될 차원수를 결정하는 것이 중요한 요소이다. 감소될 차원수는 자료의 실제 구조에 맞게 충분히 크게 선택하고 표본 추출의 오류와 중요하지 않는 상세함을 나타나지 않게 작게 선택해야 한다. 또한 감소될 차원에서 사용자의 유사성 측정을 위해 코사인 거리를 이용한다. 코사인 거리를 이용하여 유사한 사용자를 구별할 때, 코사인 거리에 대한 임계값을 결정해야 한다. 결정된 임계값에



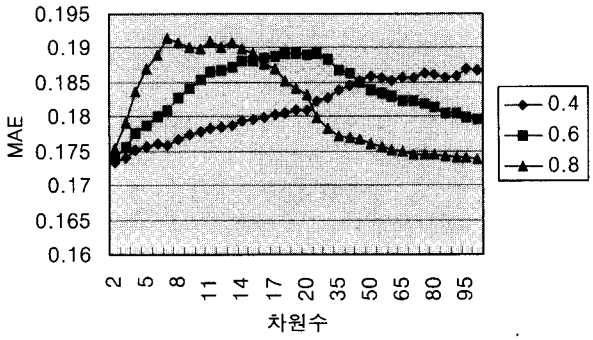
(a) 그룹 1



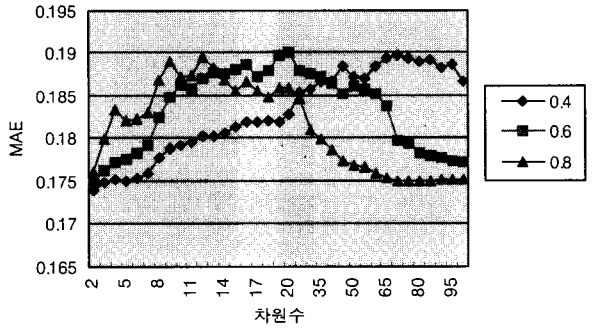
(b) 그룹 2



(c) 그룹 3

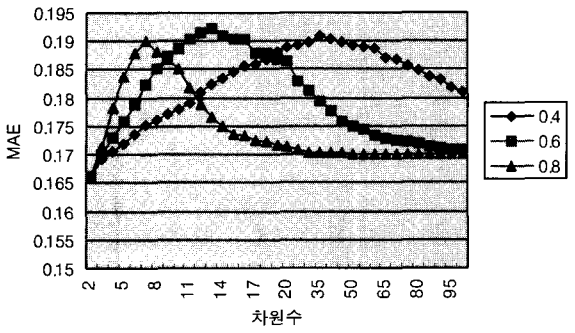


(e) 그룹 4

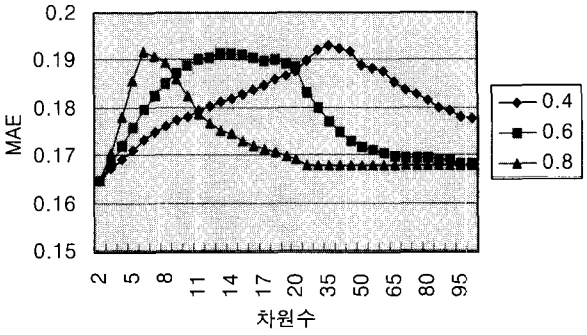


(e) 그룹 5

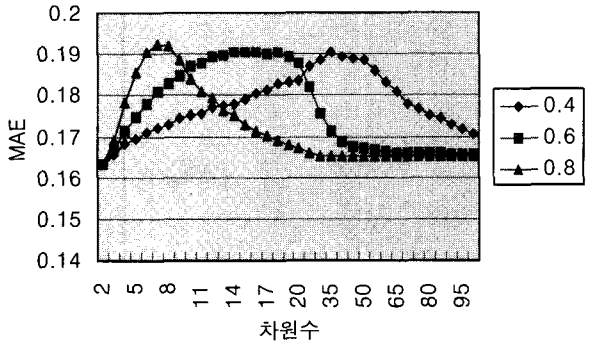
(그림 5) Eachmovie 자료에서 차원수에 따른 실험 결과



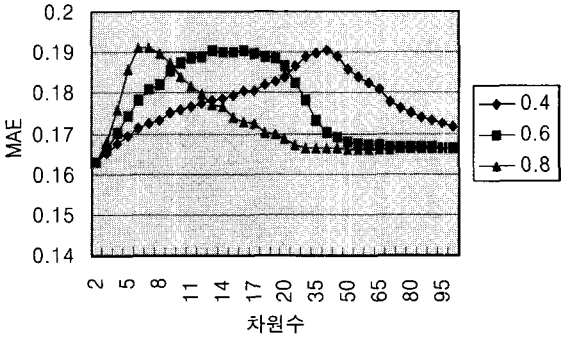
(a) 그룹 1



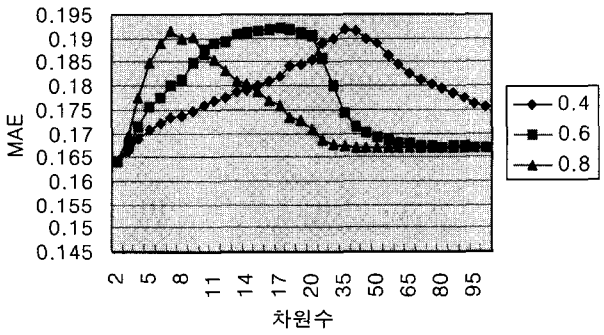
(b) 그룹 2



(c) 그룹 3



(d) 그룹 4



(e) 그룹 5

(그림 6) Movielens 자료에서 차원수에 따른 실험 결과

의해 그 값보다 큰 값을 가지는 사용자를 유사한 사용자로 구별할 수 있다.

따라서, 전체적인 실험 과정은 다음과 같다. 우선, 적절한 감소될 차원수를 결정하기 위한 실험을 수행한다. 총 10개 학습 자료를 이용하여 2에서 20까지는 1만큼씩 변경하고, 20부터 100까지는 5만큼씩 변경하면서 평가 정보를 차원수를 감소시켜서 감소된 차원에서 사용자가 평가하지 않은 것에 대한 선호도를 예측하여 검사 자료와 MAE 값을 측정한다.

두 번째로, 유사성에 대한 임계값을 결정하기 위한 실험을 수행한다. 첫 번째 실험에서 결정된 감소될 차원수로 평가 정보의 차원을 감소시킨 후에 코사인 거리를 측정한다. 코사인 거리의 임계값을 0.2에서 0.5까지는 0.1만큼씩 변경하고, 0.5에서 1.0까지는 0.05씩 변경하면서 첫 번째 실험과 같이 MAE를 측정한다.

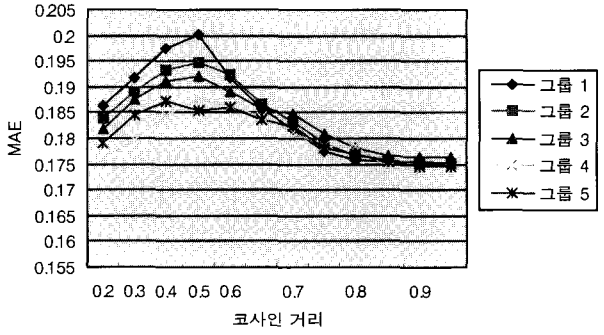
세 번째로, 감소된 차원에서 협력적 여과의 성능을 비교 평가하기 위하여 상관계수를 이용한 방법과 비교 평가한다.

첫 번째 실험 결과는 (그림 5)와 (그림 6)에서 보여진다. Eachmovie에서는 코사인 거리의 임계값이 0.8일 때, 약 50 차원부터 MAE 값이 점근적으로 접근하는 것을 관찰할 수 있었다. 코사인 거리의 임계값이 0.6일 때도 약 70차원 이상을 지나면 MAE 값이 점근적으로 접근하는 것을 알 수 있었다. 임계값이 0.4일 때 MAE 값이 수렴하는 특성을 보이지 않았다. 실험 결과로부터, 감소될 차원수가 너무 작아지면 감소된 차원에서 원래 평가 정보가 가지고 있는 특성을 잘 반영하지 못하는 것을 알 수 있었다. 또한 감소될 차

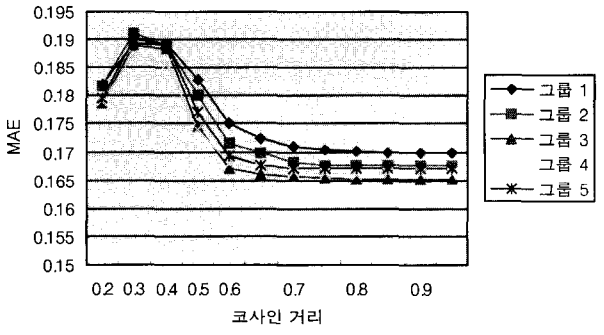
원수가 일정 차원이상 커진다고 해도 평가 정보의 특성을 더 잘 표현하지 못하는 것을 알 수 있었다.

Movielens에서도 코사인 거리의 임계값이 0.8일 때, 대략 20차원에서부터 일정한 값에 점근적으로 접근하는 것을 관찰할 수 있었다. 0.6일 때로, 대략 50차원 정도부터 점근적으로 접근하는 것을 알 수 있다. 0.4일 때는 MAE 값이 수렴하는 특성을 보이지 않았다. 두 자료에서 수렴 특성을 가지는 차원수의 차이는 각각 자료에서 평가수에서 기인한 것으로 예상되었다. Eachmovie 자료는 한 그룹 당 평균적으로 약 30,000개 정도의 평가 수를 가지고 있고, Movioelens 자료는 100,000개의 평가 수를 가지고 있다. 따라서, 감소될 차원은 평가 개수와 관련이 되어 지는 것을 알 수 있으며, 자료의 개수가 증가하면, 감소될 차원수는 작아져도 충분히 정확한 성능을 나타나는 것을 알 수 있었다.

코사인 거리에 관련된 실험 결과는 (그림 7)과 (그림 8)에서 보여진다. 사용자의 유사성 정도에 임계값으로서 코사인 거리 값을 구하여 위하여, 코사인 거리를 0.2에서 0.6까지는 0.1씩 변경하고, 0.6에서 0.95까지는 0.05씩 값을 변경하면서 실험을 수행하였다. 또한 감소될 차원으로서는 50차원을 사용하였다. Eachmovie 자료에서는 코사인 거리로 0.9 정도에서 가장 좋은 성능을 나타내었고, Movielens자료에는 0.7정도에서부터 좋은 성능을 나타내었다. 두 자료에서 코사인 거리의 차이는 감소될 차원수와 같이 평가 개수의 차이로 기인한 것으로 예상된다.



(그림 7) Eachmovie 자료에서 코사인 거리



(그림 8) Movielens 자료에서 코사인 거리

본 논문에서 기술하고 있는 협력적 여과 방법과 대체적으로 가장 성능이 좋다고 알려진 순수한 협력적 여과 방법 [4, 5]과 성능 비교를 위한 실험 결과는 <표 2>에서 보여진다. 실험 결과를 살펴보면, 두 자료에 대해서 순수한 협력적 여과 방법보다 전체적으로 좋은 성능을 나타내었다. 이러한 실험 결과는 평가 정보의 차원을 감소시켜서 협력적 여과를 수행하는 방법이 평가 정보가 가지는 원래 차원에서 협력적 여과를 수행하는 것만큼 정확하다는 것을 보여준다. 또한 원래 평가 정보의 차원보다 저차원에서 협력적 여과를 수행함으로 계산 비용도 감소시킬 수 있다.

〈표 2〉 상관 계수법과 비교 실험 결과

		상관 계수법	SVD
EM	그룹 1	0.1758	0.1754
	그룹 2	0.1793	0.1758
	그룹 3	0.1797	0.1768
	그룹 4	0.1758	0.1742
	그룹 5	0.1787	0.1756
	평균	**0.1779**	**0.1756**
ML	그룹 1	0.1682	0.17
	그룹 2	0.1664	0.1677
	그룹 3	0.1676	0.1653
	그룹 4	0.1682	0.1665
	그룹 5	0.1677	0.167
	평 균	**0.1676**	**0.1672**
총 평균		**0.1728**	**0.1714**

EM : Eachmovie ML : Movielens

5. 결 론

본 논문에서는 SVD를 이용한 협력적 여과 방법에 대해서 기술하였다. 기술된 방법은 사용자가 입력한 평가 자료의 차원을 감소시켜서 희소성을 최소화하는 방법이며, 사용자간의 유사성은 감소된 저차원에서 측정된다. 본 방법의 효과를 검증하기 위하여 대표적인 두 가지 실험자료인 Eachmovie 자료와 Movielens 자료를 이용하여 실험하였다. 실험적으로 감소될 차원수는 너무 작은 차원에서는 정확성이 감소되므로 적당히 충분한 차원수를 사용해야 한다는 것을 관찰할 수 있었다. 또한 감소될 차원수는 자료의 특성, 즉 자료의 개수에 영향을 받고 있는 것을 알 수 있었으며, 정확한 상관관계는 향후 과제로 연구할 필요하다. 사용자의 유사성의 임계값도 감소될 차원수를 결정하는데 중요한 요소로서 작용하였다. 유사성 임계값이 너무 작으면 보다 큰 차원수를 필요로 하였다. 따라서 유사성의 척도로 충분히 큰 값을 사용해야 하는 것을 관찰 할 수 있었다. 실험적으로 평가 정보의 차원을 감소시켜서 협력적 여과를 수행하는 방법이 평가 정보가 가지는 원래 차원에서 협력적 여과를 수행하는 것만큼 정확하다는 것을 알 수 있었다.

향후 연구 과제로는 위에서 기술한 자료의 수와 감소될

차원수와의 상관관계와 더불어, PCA, DCT 등과 같은 다른 차원 감소 방법들과 비교 평가가 필요하다.

참 고 문 헌

[1] Belkin, N. J. and Croft, B. W., "Information filtering and information retrieval-two sides of the same coin," CACM. 35(2), December, 1992.

[2] DEC Eachmovie collaborative filtering data set, http://www.research.digital.com/SRC/eachmovie/.

[3] David Goldberg, David Nichols, Brian M. Oki and Douglas Terry, Using collaborative filtering to weave an information tapestry, Communication of the ACM, 35(12), pp.61-70, December, 1992.

[4] J. Herlocker, J. Konstan, A. Borchers and J. Riedl., An algorithmic framework for performing collaborative filtering, Proceedings of the ACM Conference on Research and Development in Information Retrieval (SIGIR '99). Berkeley, CA.

[5] John S. Breese, David Hecherman and Carl Kadie., Empirical Analysis of Predictive Algorithms for Collaborative Filtering, In Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence(UAI-98), pp.43-52, San Francisco, July, 1998.

[6] Movielens data set, http://www.cs.umn.edu/Research/GroupLens/.

[7] Pattie Maes, Agents that reduce work and information overload, Communication of the ACM, Vol.37, No7, pp.31-40, 1994.

[8] Paul, R. Neophytos, I. Mitesh, S. Peter, B. John, R. GroupLens : an open architecture for collaborative filtering of netnews, In Proceedings of ACM CSCW '94 Conferece on Computer Supported Cooperative Work, pp.175-186, 1994.

[9] Resnick, P. and Varian, H. R., Recommender systems, CACM. 40(3), pp.56-58, March, 1997.

[10] Schafer, J. B., Konstan, J. A. and Riedl, J., Recommender Systems in E-Commerce, In ACM Conference on Electronic Commerce(EC-99), pp.158-166.

[11] Shardanand, Upendra., Social Information Filtering for Music Recommendation, S.M. Thesis, Program in Media Arts and Sciences, Massachusetts Institute of Technology, 1994.

[12] Steven J. Leon, Linear Algebra with Applications second edition, Macmillan publishing company, 1996.

정 준

e-mail : jjeong@im.inha.ac.kr

1999년 인하대학교 전자계산공학과(학사)

2001년 인하대학교 대학원 전자계산 공학과 (공학석사)

2001년~현재 인하대학교 대학원 전자계산 공학과 박사과정 재학중

관심분야 : 협력적 여과, 패턴 인식, 컴퓨터 비젼

이 필 규

e-mail : pkrhee@inha.ac.kr

1975년~1982년 서울대학교 전기공학과 (학사)

1982년~1985년 KIST 시스템구소데이타 통신실 연구원

1985년~1986년 East Texas State University 전산학 석사

1987년~1990년 University of SW Louisiana 전산학 박사

1991년~1992년 한국전자통신연구소 컴퓨터 연구단 선임연구원

1993년~1994년 IBM T.J. Watson Research Center 객원연구원

1992년~2001년 인하대학교 전자계산공학과 부교수

2001년~현재 인하대학교 전자계산학과 교수

관심분야 : 컴퓨터 인식, 패턴 인식, 생체 인식, 영상처리, 지능형 시스템