저차원 선형 모델을 이용한 하이브리드 협력적 여과

(A Hybrid Collaborative Filtering Using a Low-dimensional Linear Model)

고 수 정 †

(Su-Jeong Ko)

요 약 협력적 여과는 특별한 아이템에 대한 사용자의 선호도를 예측하는 데 사용하는 기술이다. 이러한 협력적 여과 기술은 사용자 기반 접근 방식과 아이템 기반 접근 방식으로 구분할 수 있으며, 많은 상업적인 추천 시스템에서 광범위하게 사용되고 있다. 본 논문에서는 저차원 선형 모델을 사용하여 사용자 기반과 아이템 기반을 통합하는 하이브리드 협력적 여과 방법을 제안한다. 제안한 방법에서는 저차원 선형 모델 중 비음수 행렬 분해(NMF)를 이용하여 기존의 협력적 여과 시스템의 문제점인 희박성과 대용량성의 문제점을 해결한다. 협력적 여과 시스템에서 NMF를 이용하는 방법은 사용자를 의미 관계로 표현할 때 유용하게 사용되나 사용자-아이템 행렬의 평가값에 따라 정확도가 낮아질 수 있으며, 모델 기반의 방법이기 때문에 계산 과정이 복잡하여 동적인 추천이 불가능하다는 단점을 갖는다. 이러한 단점을 보완하기 위하여 제안된 방법에서는 NMF에 의해 군집된 그룹을 대상으로 TF-IDF를 이용하여 그룹의 특징을 추출한다. 또한, 아이템 기반에서 아이템간의 유사도를 계산하기 위하여 상호정보량(mutual information)을 이용한다. 오프라인 상에서 훈련집합의 사용자를 군집시키고 그룹의 특징을 추출한 후, 온라인 상에서 추출한 그룹의 특징을 이용하여 새로운 사용자를 가장 최적의 그룹으로 분류함으로써 사용자를 분류하는 데 걸리는 시간을 단축시켜 동적인 추천을 가능하게 하며, 사용자 기반과 아이템 기반을 병합함으로써 기존의 방법보다 정확도를 높인다.

키워드 : 사용자 기반과 아이템 기반의 협력적 여과, NMF, 저차원 선형 모델, TF-IDF, 상호정보량

Abstract Collaborative filtering is a technique used to predict whether a particular user will like a particular item. User-based or item-based collaborative techniques have been used extensively in many commercial recommender systems. In this paper, a hybrid collaborative filtering method that combines user-based and item-based methods using a low-dimensional linear model is proposed. The proposed method solves the problems of sparsity and a large database by using NMF among the low-dimensional linear models. In collaborative filtering systems the methods using the NMF are useful in expressing users as semantic relations. However, they are model-based methods and the process of computation is complex, so they can not recommend items dynamically. In order to complement the shortcomings, the proposed method clusters users into groups by using NMF and selects features of groups by using TF-IDF. Mutual information is then used to compute similarities between items. The proposed method clusters users into groups and extracts features of groups on offline and determines the most suitable group for an active user using the features of groups on online. Finally, the proposed method reduces the time required to classify an active user into a group and outperforms previous methods by combining user-based and item-based collaborative filtering methods.

Key words : User-based and item-based collaborative filtering, NMF, a Low-dimensional linear model, TF-IDF, mutual information

· 이 연구는 인덕대학 학술연구비 일부 지원에 의하여 수행되었음

† 종신회원 : 인덕대학 컴퓨터소프트웨어과 교수
sjko@induk.ac.kr

논문접수 : 2009년 3월 5일
심사완료 : 2009년 8월 12일

정보과학회논문지 : 소프트웨어 및 응용 제36권 제10호(2009.10)

1. 서 론

협력적 여과는 사용자가 아이템에 대해 평가한 흥미의 정도를 기반으로 사용자의 아이템에 대한 선호도를 예측함에 목표를 두고 있다. 이러한 기술은 사용자 기반 접근과 아이템 기반 접근 방식으로 나눌 수 있다[1]. 사용자 기반의 협력적 여과는 추천 시스템에서 가장 성공적인 시스템이며 상업적으로도 광범위하게 사용되고 있다[2]. 일반적으로, 사용자 기반의 협력적 여과 시스템은 가장 높은 가중치를 갖는 N개의 아이템을 선택하고 이를 사용자에게 추천하는 Top-N 추천 기법을 사용한다[3]. 사용자 기반의 추천 시스템이 상업적으로 성공적이었을지라도 시스템의 대용량성과 실시간 처리에서의 문제점으로 인해 한계점을 갖는다[4]. 이러한 한계점을 보완하기 위하여, 사용자 기반 방법과 아이템 기반 방법을 병합하는 방법을 사용한다[1,5]. 예를 들어, 사용자 기반 추천 시스템에서 가장 유사도가 높은 이웃을 선정하는 과정에서의 복잡도를 보완하는 방법으로 아이템을 군집시키는 방법이 사용된다[6].

본 논문에서는 저차원 선형 인수 모델을 사용하여 사용자 기반과 아이템 기반을 병합하는 협력적 여과를 제안한다. 기존의 협력적 여과 시스템에서 사용자-아이템 행렬에 선형 모델을 적용하여 저차원으로 감소시키는 방법은 매우 효과적인 방법임이 증명되었다[7,8]. 특히, 저차원 선형 모델 중 비음수 행렬 분해(NMF, Non-negative Matrix Factorization)를 이용한 방법은 높은 정확도를 나타내고, 협력적 여과 시스템의 문제점인 희박성 문제를 해결할 수 있음을 보였다[9]. NMF는 단지 최소한의 인수만으로 사용자의 선호도에 영향을 줄 수 있으며, 각 인수가 사용자와 아이템에 어떻게 적용되는지를 결정함에 의해 아이템에 대한 사용자의 선호도가 결정된다[10]. 또한, 대용량의 데이터베이스에서 지수적인 함수 관계를 발견하는 데 있어서 강력한 알고리즘이다. NMF를 이용한 방법은 저차원 선형 모델의 하나인 SVD(Singular Value Decomposition)와 같은 기존의 방법과 달리 직교적이지 않은 사용자간일지라도 사용자들간에 의미 관계를 발견할 수 있으며, 또한 비음수적으로 분해를 하기 때문에 사용자간의 의미 관계를 정확도 높게 표현한다[11-13]. 반면, NMF만을 이용한 협력적 여과 방법은 불완전한 평가값을 대상으로 모델 기반에서 결측치를 예측하는 방법이기 때문에 계산과정이 복잡하고 시간이 많이 소요되어 실시간에서의 동적인 추천이 어렵다는 단점을 갖는다[14]. [8]에서는 평가값이 완전하지 않음으로 인해 발생하는 정확도 저하의 문제점을 해결하기 위하여 EM알고리즘을 이용하여 잡음을 제거한 후에 가중치가 부여된 NMF를 사용하는 방법을 사용하였다. 이와 같은 방법은 NMF만을 사용한 방법보다는 정확도가 높으나 초기값에 민감하고 혼합성분의 개수를 미리 알고 있어야 한다는 단점과 모델 기반에서 계산하므로 동적인 추천이 불가능하다는 단점을 갖는다. NMF 를 응용한 직교의 비음수 행렬 3-분해(ONMTF, Orthogonal Nonnegative Matrix Tri-factorization)를 이용한 방법은 모델 기반의 방법과 메모리 기반의 단점을 보완하고 각 방법의 장점을 병합하는 방법이다[9]. ONMTF를 이용한 방법은 사용자-아이템 행렬을 두 개의 인수가 아니고 3개의 인수로 분해하며, 분해된 3개의 인수를 이용해 사용자-아이템 군집의 중심을 찾으며, 이 중심과 새로운 사용자 사이의 유사도를 계산한다. 다음으로, 가장 유사도가 높은 그룹을 찾고 사용자와 아이템의 이웃을 선정한다. 메모리 기반의 접근 방식을 이용하여 새로운 사용자의 결측치를 예측하고, 코사인 유사도를 이용하여 아이템에 대한 결측치를 예측한다. 마지막으로, ONMTF, 사용자 기반의 예측 방법, 그리고 아이템 기반의 예측 방법에 의한 값들을 혼합 계수에 의해 선형적으로 병합하여 최종적으로 예측값을 결정한다. ONMTF는 메모리 기반과 모델 기반의 접근 방식을 병합함으로써 정확도를 높일 수 있었으나 혼합 계수에 따라 정확도가 일정하지 않고, 혼합 계수도 자동적으로 결정할 수 없다는 단점을 갖는다. [14]의 연구에서는 SVD나 NMF보다는 공군집화(co-clustering)를 사용하고, 실시간 처리를 위해 병렬 처리 방법을 제안하였다. 이 방법은 모델기반의 방법이므로 메모리 기반의 방법보다는 정확도가 낮다는 단점이 있으며, 또한 병렬 처리를 위한 시스템이 갖춰져야 한다는 단점도 갖는다.

본 논문에서 제안한 방법에서는 기존의 선형 인수 모델을 사용한 협력적 여과 방법의 단점을 보완하기 위하여 NMF을 사용하여 사용자간의 의미 관계를 표현한 후, 군집된 그룹의 특징을 추출하기 위하여 TF-IDF를 사용한다. 새로운 사용자나 아이템을 추가하기 위하여 전체 모델을 다시 계산하지 않고 TF-IDF를 이용하여 추출한 그룹의 특징을 이용하여 새로운 사용자나 아이템을 동적으로 분류함으로써 기존 모델 기반 알고리즘의 단점을 보완한다. 또한, 사용자 정보에 아이템 정보를 병합함으로써 사용자 기반으로만 아이템을 추천했을 경우의 단점을 보완하기 위하여 아이템간의 유사도를 계산한다. 이를 위하여 상호정보량(mutual information)을 이용하며, 이를 기반으로 아이템을 추천함으로써 추천의 정확도를 높인다.

2. 사용자 군집과 그룹의 특징 추출

본 장에서는 사용자 기반에서 사용자를 군집하고 그룹의 특징을 추출하는 방법을 기술한다. 사용자 군집을

위하여 LSI(Latent Semantic Indexing) 방법들 중 NMF를 사용하여 사용자들을 의미 속성(semantic attribute)으로 표현한다. 또한, 그룹의 특징을 추출하기 위하여 TF-IDF를 사용하는 방법을 기술한다.

2.1 NMF를 사용한 사용자 군집

본 논문에서 제안한 방법은 협력적 여과 추천을 위하여 아이템-사용자 행렬을 사용한다. 아이템-사용자 행렬은 CI × CU의 행렬 M으로 정의한다. 또한, 벡터 k는 $\vec{k}$로 정의한다. 행렬 M은 n × m의 차원으로 정의되며, 행렬의 요소는 m_{ij}다. 행렬 M은 비음수 행렬이며, 식 (1)과 같이 K와 Y를 사용하여 정의된다[11].

$$M \approx KY \quad (1)$$

NMF는 근사의 분해를 수행하며 다변량적인 자료의 통계 분석에 적용된다. m차원 벡터가 주어진다면 그 벡터는 행렬 M의 열과 1대 1로 대응된다. 이는 데이터 집합의 사용자 수 m를 표현한다. 행렬 M은 행렬 K와 행렬 Y로 분해되고, 행렬 K의 기저 벡터 집합과 행렬 Y의 히든 벡터 집합을 찾는다. 행렬 Y의 열은 행렬 M의 열과 일대일 대응관계를 찾으며, 그 결과는 행렬 K, 기저벡터의 가중치 합이다. 비음수 행렬 분해를 위해 $\|M-KY\|^2$는 $K \geq 0$와 $Y \geq 0$로 최소화된다[12]. 행렬 M과 분해된 두 행렬 K와 Y가 모두 비음수 행렬, 즉 행렬의 모든 요소가 0 이상의 값을 갖는다. 결과적으로, 비음수 행렬 분해는 두 행렬 K와 Y의 곱 KY와 행렬 M 사이의 차이를 최소화하는 비음수 행렬 K와 Y를 찾는 것이 목적이며, 이때 행렬 K는 n × r로 Y는 r × m의 행렬로 분해된다. 여기서, r은 n와 m보다 작다.

M은 r개의 의미 속성을 갖는 K와 Y로 분해되며, 식 (2)와 같이 표현된다.

$$M = K' \cdot Y'^T = [\vec{k}_1,..,\vec{k}_a,..,\vec{k}_r][\vec{y}_1,..,\vec{y}_a,..,\vec{y}_r]^T \quad (2)$$

식 (2)에서 K'은 n × r의 비음수 행렬이며 Y'는 m × r의 비음수 행렬이다. 열벡터 $\vec{k_a}$는 $(k_{a1},k_{a2},\cdots,k_{aj},\cdots,k_{an})$의 요소이며, $\vec{y_a}$는 $(y_{a1},y_{a2},\cdots,y_{ai},\cdots,y_{am})$의 요소이다. $\vec{k_a}$와 $\vec{y_a}^T$의 열벡터는 각각 r개의 의미 속성 중에서 a번째 속성의 사용자와 아이템에 대한 가중치 벡터이다.

표 1은 행렬 M을 비음수 인자로 분해하기 위한 아이템-사용자 행렬의 예이다. 표 1은 20명의 사용자와 13개의 아이템으로 구성되었다. 50명의 사용자와 30개의 아이템을 GroupLens 추천 시스템[15]으로부터 무작위로 추출한다. 추출한 집합을 대상으로 결측치가 적은 수로 정렬하고, 정렬된 값 중 상위의 값을 나타내는 데이터를 선정하여 20명의 사용자와 13개의 아이템으로 표 1을 구성한다. 아이템-사용자 행렬의 결측치에 대한 기본 평가값은 평가된 모든 값의 평균값으로 결정하고, 결측치를 그 평균값으로 대체하였다.

그림 1은 표 1의 아이템-사용자 행렬을 식 (1)의 형태와 같이 비음수 행렬로 분해를 한 결과이다.

표 1 아이템-사용자 행렬 M의 예

	U_1	U_2	U_3	U_4	U_5	U_6	U_7	…	U_{12}	U_{13}	U_{14}	U_{15}	U_{16}	U_{17}	U_{18}	U_{19}	U_{20}
i_1	1	0.8	1	0.8	1	0.2	*0.7*	…	1	0.8	0.8	1	1	0.6	*0.7*	0	0.2
i_2	0.8	0.4	0.8	*0.7*	1	0	0.6	…	0.6	0.8	*0.7*	0.6	0.6	0	0.4	0.6	0.6
i_3	0.8	0.4	0.8	*0.7*	0.8	0.4	0.6	…	0.8	0.8	0.8	0.4	0.8	0.6	0.8	0.6	0.6
i_4	1	0.8	0.8	1	1	1	0.2	…	*0.7*	1	*0.7*	*0.7*	*0.7*	1	0.8	*0.7*	*0.7*
…	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
i_{11}	0.8	0	0	0.8	1	0	0	…	*0.7*	*0.7*	0.8	*0.7*	0.6	0.8	0.8	*0.7*	0
i_{12}	1	*0.7*	1	1	1	0	0.6	…	0.7	1	0.6	0.8	0.6	0.4	1	0.8	*0.7*
i_{13}	0.8	0.6	1	*0.7*	1	*0.7*	0.6	…	0.8	*0.7*	0.8	*0.7*	*0.7*	*0.7*	*0.7*	0.4	*0.7*

Factor K=

0.756	0.004	0.878	0.000	0.283
0.414	0.334	0.481	0.407	0.000
0.469	0.220	0.357	0.356	0.438
0.315	0.487	0.216	0.432	0.790
0.244	0.002	0.492	0.482	0.645
0.833	0.000	0.000	0.126	0.000
0.349	0.099	0.613	0.462	0.526
0.089	0.586	0.526	0.634	0.000
0.433	0.034	0.000	0.551	0.617
0.341	0.264	0.729	0.605	0.165
0.255	1.281	0.000	0.000	0.095
0.298	0.659	0.551	0.521	0.000
0.491	0.000	0.536	0.294	0.631

Factor Y=

0.789	0.00	0.623	0.558	0.226	0.00	0.00	1.054	0.698	0.000	0.138	0.520	0.651	0.663	0.000	0.436	0.000	0.132	0.047	0.092
0.549	0.00	0.000	0.498	0.663	0.00	0.00	0.507	0.578	0.728	0.458	0.394	0.447	0.364	0.518	0.328	0.515	0.598	0.481	0.000
0.328	0.69	0.542	0.272	0.771	0.00	0.92	0.207	0.410	0.671	0.521	0.507	0.270	0.307	1.017	0.472	0.260	0.653	0.000	0.146
0.294	0.48	0.608	0.690	0.055	0.00	0.44	0.383	0.039	0.193	0.274	0.087	0.666	0.342	0.000	0.000	0.000	0.068	0.876	1.136
0.150	0.42	0.108	0.219	0.615	1.03	0.00	0.065	0.306	0.000	0.423	0.379	0.181	0.325	0.182	0.445	0.947	0.377	0.221	0.201

그림 1 NMF를 사용한 행렬 M의 분해

사용자 기반의 협력적 여과 방식이므로 아이템은 제외하고 사용자를 기반으로 의미 속성을 선택한다. 식 (3)에서 행렬 M은 r개의 의미 속성을 가진 사용자와 아이템 사이의 관계로서 표현된다. 이 식에서 아이템의 가중치는 행렬 K'의 각 열벡터의 크기(norm)를 1로 정규화되도록 제거된다.

$$K'=\left[\frac{\vec{k}_1}{\|\vec{k}_1\|_2},...,\frac{\vec{k}_r}{\|\vec{k}_r\|_2}\right], Y'=\left[\vec{y}_1\cdot\|\vec{k}_1\|_2,...,\vec{y}_r\cdot\|\vec{k}_r\|_2\right] \quad (3)$$

식 (3)에서 $\|\vec{k}_r\|_2$는 벡터 $\vec{k}_r$의 유클리드 길이를 나타내며 식 (4)와 같이 정의한다.

$$\|\vec{k}_r\|_2=(k_{a1}{}^2,k_{a2}{}^2,...,k_{an}{}^2)^{1/2} \quad (4)$$

표 2는 그림 1의 행렬 K'와 행렬 Y'에 식 (3)을 적용하여 행렬 K'의 각 열벡터의 크기를 1로 정규화하고, 행렬 K'의 가중치를 식 (3)을 이용하여 행렬 Y'에 부여한 후, Y'를 5개의 의미 속성으로 표현한 결과이다.

표 2를 기반으로 보면 20명의 사용자는 5개의 속성으로 분류된다. 그리고 각 사용자는 5개의 그룹 중 하나의 클래스로 분류할 수 있다. 예를 들어, 표 2에서 U_1은 {그룹1, 그룹2, 그룹3, 그룹4, 그룹5} 중에서 그룹1에 가장 큰 가중치를 나타내므로, 그룹1로 분류한다. 이와 같은 원리로 표 1의 사용자는 그림 2의 결과로 분류된다.

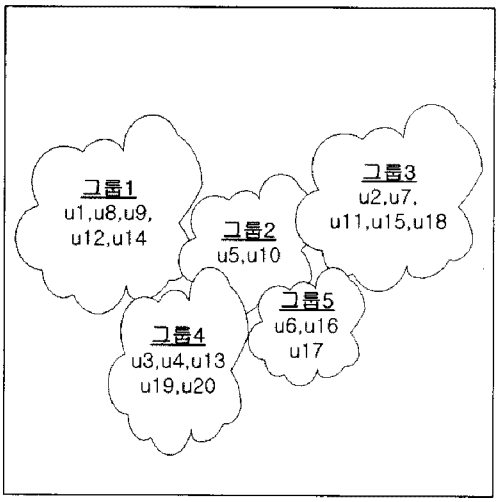

그림 2 그룹으로 분류된 사용자

2.2 그룹의 특징 선택

베이지안 네트워크 또는 SVD를 기반으로 하고 있는 모델 기반의 알고리즘은 새로운 사용자나 아이템을 추가할 경우, 유사도 계산을 위해 전체 행렬을 다시 계산하여 시간적인 소모가 크다는 단점을 갖는다[2]. 제안된 방법은 전체 행렬을 다시 계산하지 않고 새로운 사용자와 군집된 그룹의 특징만을 비교함으로써 기존 사용자 기반 협력적 여과 시스템의 단점을 해결한다.

그룹의 특징을 추출하기 위해 사용자들이 그룹 내에서 높게 평가한 아이템들을 선정하고, 그 아이템들이 그룹 내에서 어떠한 빈도로 추출되었는가를 계산한다. 반면, 빈도만을 기반으로 그룹의 특징을 선택하였을 경우, 인기가 있는 아이템의 평가값 대부분이 사용자들에게 높은 값을 나타내므로 그룹의 특징이 인기가 있는 아이템들만으로 구성되는 한계점을 갖는다. 이러한 문제를 해결하기 위하여 제안된 방법은 TF-IDF를 사용한다[16]. TF-IDF는 정보검색분야에서 문서의 특징을 추출하는 용도로 사용되는 방법이다. 식 (5)는 그룹 k에 속한 j번째 아이템 i_{kj}의 TF-IDF를 계산하는 식이다. TF-IDF를 계산하기 위한 전처리로, 사용자들이 아이템에 대해 흥미가 있다고 판단하기 어려운 아이템, 즉 모든 평가값의 평균보다 낮은 평가값을 갖는 아이템은 아이템-사용자 행렬로부터 제외시킨다.

$$W_{nk}=f_{nk}\bullet[\log_2\frac{p}{DF}+1] \quad (5)$$

식 (5)에서 f_{nk}는 아이템 i_{kj}의 상대 빈도이며, 이는 그룹 안에 속한 모든 아이템의 평가값의 평균보다 높게 평가된 빈도를 나타낸다. p는 모든 그룹의 수이며, DF는 아이템 i_{kj}가 평균값보다 높게 평가된 그룹의 수이다. 다음으로, 식 (5)와 같이 계산한 TF-IDF를 내림차순으로 정렬하여 높은 값을 갖는 아이템을 추출한다.

표 3은 그림 2에 나타난 그룹 사용자들을 기반으로 식 (5)를 사용하여 그룹에 속한 아이템들의 TF-IDF를 계산한 결과이다.

그룹의 특징은 TF-IDF를 가중치로 하여, 벡터의 형태로 표현된다. 예를 들어 그룹1의 특징은 {0.12, 0.07, 0.12, 0.1, 0.08, 0.16, 0.1, 0.05, 0, 0.1, 0.1, 0.07, 0.13} 벡터의 13차원이다.

표 2 사용자를 의미 속성으로 표현

	U_1	U_2	U_3	U_4	U_5	U_6	U_7	U_8	U_9	U_{10}	U_{11}	U_{12}	U_{13}	U_{14}	U_{15}	U_{16}	U_{17}	U_{18}	U_{19}	U_{20}
그룹 1	0.485	0.000	0.383	0.343	0.139	0.000	0.000	0.648	0.429	0.000	0.085	0.320	0.400	0.408	0.000	0.268	0.000	0.081	0.029	0.05
그룹 2	0.323	0.000	0.000	0.293	0.400	0.000	0.000	0.298	0.340	0.428	0.269	0.232	0.263	0.214	0.304	0.193	0.303	0.351	0.283	0.00
그룹 3	0.184	0.387	0.303	0.152	0.431	0.000	0.518	0.116	0.229	0.376	0.292	0.284	0.151	0.172	0.569	0.264	0.145	0.365	0.000	0.08
그룹 4	0.191	0.316	0.395	0.448	0.036	0.000	0.288	0.249	0.025	0.125	0.178	0.057	0.433	0.222	0.000	0.000	0.000	0.044	0.569	0.73
그룹 5	0.097	0.270	0.070	0.141	0.387	0.662	0.000	0.042	0.197	0.000	0.273	0.244	0.117	0.210	0.117	0.287	0.611	0.243	0.143	0.13

표 3 그룹에 속한 아이템들의 특징 선택

	그룹1	그룹2	그룹3	그룹4	그룹5
i_1	0.12	0.07	0.07	0.1	0.14
i_2	0.07	0.07	0.1	0.07	0.07
i_3	0.12	0.07	0.1	0.07	0.14
i_4	0.1	0.14	0.07	0.1	0.14
...	..	..	..	..	..
i_{11}	0.1	0.14	0.03	0.03	0.14
i_{12}	0.07	0.14	0.1	0.13	0.07
i_{13}	0.13	0.09	0.09	0.04	0

3. 아이템 유사도 측정 및 추천

본 장에서는 상호정보량을 사용하여 아이템간의 유사도를 계산하는 방법과 사용자 기반과 아이템 기반을 병합하여 아이템을 추천하는 방법이 기술된다.

3.1 아이템간의 유사도 계산

계수 측정 또는 연관도 측정은 자연어 처리 분야에서 단어의 동시 발생 빈도를 계산하는 데 사용되었다. 코사인 계수 등 다양한 연관도 측정 척도가 자연어 처리에서 사용되어왔다. 본 논문에서 사용된 방법에서는 자연어 처리의 '단어'를 '아이템'으로 간주하고, 상호정보량을 이용하여 유사도를 계산한다. 상호정보량은 Churck와 Hanks[17]가 이를 Shannon의 정보이론[18]을 기반으로 하는 수집분석[19]에 적용한 이후 많은 분야에서 사용되어 왔다. 상호정보량은 사건 x와 사건 y, 각각의 독립적인 발생 확률에 대해 사건 x와 사건 y가 동시에 발생할 확률을 평가한다. 본 논문에서 제안된 방법에서는 사건 x를 아이템 i_x의 빈도로, 사건 y는 아이템 i_y의 빈도로 정의한다. 이와 같이 정의된 아이템 i_x와 아이템 i_y의 상호정보량은 식 (6)으로 정의한다.

$$MI(x,y) = \log_2 \frac{p(x,y)}{p(x)p(y)} \tag{6}$$

식 (6)에 나타난 p(x)와 p(y), 그리고 p(x,y)는 식 (7)과 식 (8)로 각각 정의된다. p(x)와 p(y)는 모든 아이템 수(n)에 대해 사용자가 아이템 i_x와 아이템 i_y를 평가한 각각의 빈도를 나타낸다. p(x,y)는 모든 아이템의 수(n)에 대해 사용자가 아이템 i_x와 아이템 i_y를 동시에 평가한 빈도를 나타낸다.

$$p(x) = \frac{f(x)}{n}, p(y) = \frac{f(y)}{n} \tag{7}$$

$$p(x,y) = \frac{f(x,y)}{n} \tag{8}$$

또한, f(x,y)가 0일 경우, 즉 동시에 평가된 아이템이 없는 경우는 식 (9)를 사용한다.

$$MI(x,y) = -1 \qquad if \ldots p(x,y) \neq 0 \tag{9}$$

아이템간의 유사도를 정확하게 계산하기 위해서 사용자가 아이템에 대해 흥미롭다고 평가한 값만을 유사도 계산의 대상으로 한다. 이를 위해, 표 1의 평가값을 대상으로 전체 평가값의 평균을 계산하고, 평균보다 높은 평가값을 갖는 아이템만을 흥미가 있는 아이템으로 간주한다. 표 4는 식 (6)과 식 (9)를 사용하여 아이템간의 상호정보량을 계산한 결과를 나타낸다.

표 4에서 아이템들간의 상호정보량은 0보다 작은 값들도 포함되기 때문에 이를 유사도로 사용하여 결측치 예측에 사용하기 어렵다. 따라서, 이들을 0보다 큰 값으로 식 (10)을 이용하여 대체시킨다. 다음으로, 아이템간의 상대 유사도를 구하기 위해 대체된 값을 1의 크기로 정규화시킨다.

$$MI'(x,y) = MI(x,y) + \underset{k \in I}{MIN}\, MI(x,k) \tag{10}$$

식 (10)에서 MI'(x,y)는 0보다 큰 값으로 대체시킨 값이며, I는 아이템들의 집합을 나타낸다. $\underset{k \in I}{MIN}\, MI(x,k)$은

표 4 아이템간의 상호정보량

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}	i_{11}	i_{12}
i_1												
i_2	0.474											
i_3	0.222	1										
i_4	0.085	0.278	-0.14									
i_5	0.737	-0.26	0	0.447								
i_6	0.737	0.737	1	-0.14	1.322							
i_7	0.585	0.415	0	-0.14	0.585	1						
i_8	0.152	0.567	-0.17	0.015	0.152	0.152	0.415					
i_9	-0.26	-1	-1	-0.14	1.322	-1	0	0.152				
i_{10}	0.252	0.252	0	0.225	0.515	0.515	0.363	0.152	0.515			
i_{11}	0.152	0.152	0.152	0.5	0.152	1.152	0.415	-0.02	0.152	0.345		
i_{12}	0.278	0.599	0.126	0.573	0.447	-0.14	0.348	0.693	-0.14	0.377	0.5	
i_{13}	0.737	0.737	0.737	-0.14	0.737	0.737	0.415	-0.43	-1	0.252	0.567	0.278

아이템 i_x와 다른 아이템들과의 상호정보량 중에서 가장 작은 값을 구하는 식이다. 구한 값을 기존의 상호정보량에 더하여 0보다 작은 값을 0이상의 값으로 대체시킨다.

3.2 아이템 추천

전통적인 협력적 여과 기술의 사용자 군집 모델은 사용자와 유사한 사용자를 찾아서, 그 사용자가 흥미롭다고 평가한 아이템들을 추천 하는 방법을 사용하였으나 아이템 기반의 협력적 여과 기술[20,21]은 사용자가 선택한 아이템과 유사한 아이템들을 찾아 추천하는 것이다. 아이템 기반의 협력적 여과 기술의 핵심은 아이템들간의 유사도를 나타내는 아이템-아이템 행렬로서 이러한 정보는 아이템-사용자 행렬의 요소를 분석함으로써 생성할 수 있다. 이러한 과정은 복잡하고 계산량이 많으나 오프라인에서 가능하기 때문에 활용도 역시 높으며, 유사도가 높은 아이템들을 비교하므로 추천의 정확도가 높다. 협력적 여과 추천 방법의 대표 사이트인 Amazon.com역시 아이템 기반의 협력적 여과 기술을 제안하고 사용하고 있다[22].

이러한 아이템 기반의 협력적 여과 기술은 추천을 위해 Top-N의 아이템을 선택하는 방법과 사용자가 평가하지 않은 아이템에 대한 결측치를 예측하는 방법으로 구분할 수 있다[20,21]. 본 논문에서 제안한 방법에서는 새로운 사용자를 그룹으로 분류하고, 그 사용자가 평가한 아이템 중에서 가장 높은 평가값을 갖는 아이템과 다른 아이템들간의 유사도를 계산한 후에, Top-N의 아이템을 선택한다. 그리고 아이템 기반의 협력적 여과 방법에 의해 새로운 사용자가 평가하지 않은 아이템의 결측치를 예측한다.

새로운 사용자와 그룹내의 사용자의 유사도를 계산하기 위해서 2.2절에 정의된 그룹의 특징과 새로운 사용자의 평가값을 이용한다. 유사도 계산을 위해서 유사도 계산에 대표적으로 많이 사용되는 코사인 유사도[23]를 이용한다. 새로운 사용자와 그룹의 대표 평가값과의 유사도를 계산한 후에, 그 결과를 기반으로 새로운 사용자를 그룹으로 분류하고, 사용자에게 아이템을 추천한다. 예를 들어, 사용자 U_a가 표 5와 같이 평가하였을 때 사용자 U_a와 각각의 그룹, {그룹1, 그룹2, 그룹3, 그룹4, 그룹5}와의 유사도를 코사인 유사도 식에 따라 구하였을 때, 그 결과는 {0.904, 0.844, 0.744, 0.748, 0.837}이다.

표 5 사용자 U_a의 평가값

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}	i_{11}	i_{12}	i_{13}
U_a	1	0.6	0.8	0.8	0.2	1	0.8	0.6		0.6	1	0.6	

표 5의 사용자 U_a는 그룹과의 유사도가 가장 높은 그룹 1로 분류된다. 표 4의 아이템-아이템 행렬을 기반으로 i_9에 대해 유사도가 높은 순서에서 낮은 순서로 정렬을 하면 {i_5,i_{10},i_8,i_{11},i_7,i_4,i_{12},i_1,i_2,i_3,i_6,i_{13}}의 순서로 정해진다. 아이템 추천을 위해 Top-5을 이용하였을 경우, 아이템들 중에서 i_5,i_{10},i_8,i_{11},i_7이 추천된다.

반면, 아이템 i_9에 대한 결측치를 계산하고자 할 경우, 새로운 사용자가 평가한 아이템들과 i_9와의 유사도를 가중치로 정하고, 아이템 기반의 협력적 여과에 사용하는 식[21]을 사용한다. 식 (11)은 아이템 기반의 협력적 여과에서 사용하는 식으로, 사용자 U_a가 아이템 i_m에 대해 평가한 결측치($x_{a,m}$)를 예측한다.

$$x_{a,m} = \frac{\sum_{i_b \in w(i_m)} w(i_m, i_b)(x_{a,b})}{\sum_{i_b \in w(i_m)} w(i_m, i_b)} \quad (11)$$

식 (11)에서 i_b는 전체 아이템들 중 사용자 U_a가 평가한 아이템만을 선별한 아이템들의 집합이고, $w(i_m,i_b)$는 아이템 i_m과 아이템 i_b의 유사도로, 상호정보량을 정규화한 값을 이용한다. i_b는 아이템 i_m과의 유사도를 높은 순서에서 낮은 순서로 정렬하였을 경우, 유사도가 높은 Top-N의 아이템 집합이다. 식 (12)는 식 (11)의 $w(i_m)$을 정의하는 식이다.

$$w(i_m) = \{i_b \mid rankw(i_m, i_b) \le N, x_{a,b} \neq \phi\} \quad (12)$$

식 (11)을 이용하여 아이템 i_9의 결측치를 예측하기 위해서는 3.1절에 기술된 것과 같이 상호정보량을 정규화하는 과정이 선행되어야 한다. 이를 위하여, 우선 표 6과 같이 표 4에 있는 전체 상호정보량 중에서 아이템 i_9와 다른 아이템과의 유사도만을 선정하여, 식 (10)을 사용하여 대체시킨다.

표 6 대체된 i_9와 다른 아이템과의 상호정보량

i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}	i_{11}	i_{12}	i_{13}
0.737	0	0	0.862	2.322	0	1	1.152	2	1.515	1.152	0.862	0

다음으로, 표 6의 대체된 값을 정규화하면 표 7과 같이 아이템 i_9와 다른 아이템과의 유사도를 얻을 수 있다.

표 7 정규화된 i_9와 다른 아이템과의 유사도

i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}	i_{11}	i_{12}	i_{13}
0.163	0	0	0.190	0.514	0	0.221	0.255	0.442	0.335	0.255	0.190	0

표 7의 값을 식 (11)의 $w(i_m,i_b)$로 정의하고, 식 (11)에 대입할 경우, 사용자 U_a에 대한 아이템 i_9의 결측치는 0.621로 예측된다.

4. 성능 평가

본 논문에서 제안한 방법이 추천의 정확도를 향상시

켰다는 것을 증명하기 위해 기존에 있는 추천 시스템의 데이터에 기존의 성능 측정 척도를 사용하였다.

4.1 실험집합

실험을 위해, MovieLens 추천 시스템의 데이터를 사용하였다. MovieLens 시스템은 웹기반의 연구 추천 시스템이며 1997년에 만들어졌다[15]. 사용자들은 데이터베이스에 150,000개의 평가값이 획득될 때까지 무작위하게 선택되었다. 사용자들 중 아이템에 대해 20보다 적은 평가를 한 사용자는 실험 대상에서 제외하였다. 이러한 데이터베이스는 70%의 훈련 집합과 30%의 테스트 집합으로 구분하였다. 데이터 집합은 3012의 행과 5834의 열을 가진 아이템-사용자 행렬로 전환되었다.

4.2 성능 평가 척도

실험에 사용된 평가 척도는 식 (13)에 나타난 MAE (Mean Absolute Error)를 사용하였다. 이 식은 테스트 사용자가 아이템에 대해 평가한 실제 평가값과 예측 평가값의 평균에 대한 절대 표준편차를 나타낸다[24].

$$MAE = \frac{1}{n}\sum_{j=1}^{n}\left|r_{uj} - \hat{r}_{uj}\right| \tag{13}$$

MAE의 한계는 평가 상황에 관계없이 모두 같은 값으로 다룬다는 것이다. 예를 들어 MAE는 $(r_{uj}, \hat{r}_{uj})$의 값이 (0.2, 0.8)와 (0.8, 0.2)인 경우, 두 경우에 아무런 차이를 두지 않는다. 그러나 아이템의 차이점을 고려한다면 (0.2, 0.8)와 (0.8, 0.2)는 다른 상황으로 처리되어야 할 것이다[24].

또한, 제안된 방법은 테스트 집합에서 Top-N의 아이템을 추천하였다. 사용자가 선정한 아이템이 Top-N과 일치했다면 그 집합을 히트 집합이라고 정의한다. 정확도와 재현율은 기존의 정보 검색 시스템에서 성능을 평가하기 위해 사용되어 왔다. 이와 같은 정확도와 재현율을 추천 시스템에 사용하기 위해 Top-N 추천에 관계된 정의가 필요하다. 재현율은 전체 테스트 집합 사이즈에 대한 히트 집합 사이즈의 비율이다. 식 (14)에 정의된 측정인자는 정확도와 재현율의 조화평균이다.

$$F1 = \frac{2 * precision * recall}{(precision + recall)} \tag{14}$$

사용자에게 적합한 아이템이 무엇인지를 정의하는 것은 중요하다. 이에 대한 논제를 다룬 논문[24]이 있으며 이 논문에서는 0.6보다 크게 평가한 아이템을 사용자에게 적합한 아이템이라고 간주하고 있다.

4.3 실험 결과

본 논문에서 제안한 저차원 선형 모델을 사용하고 있는 병합 협력적 여과 방법(CUSL_R)은 피어슨 상관 계수를 사용하는 방법(Pearson_R)[25], 아이템 기반의 협력적 여과를 사용하고 있는 방법(SL_R)[21], 그리고 사용자 기반의 협력적 여과를 사용하고 있는 방법(USL_R)[26]과 MAE를 사용하여 비교되었다. 또한, CUSL_R은 협력적 여과에 NMF를 적용하는 기존의 방법들과 MAE를 사용하여 비교되었다. NMF를 사용하는 기존의 방법은 모델 기반에서 행렬 분해를 사용하는 방법(NMF), EM알고리즘을 이용하여 잡음을 제거한 후 NMF를 사용하는 방법(WNMF), co-clustering을 이용하는 방법(COCLUST), 그리고 NMF군집을 기반으로 모델 기반과 메모기 기반을 병합하는 방법인 ONMTF이다. 마지막으로, CUSL_R은 SVD를 사용하고 있는 방법(SVD_R)[27]과 ClustKNN_R를 사용하는 방법(ClustKNN_R)[28]과 F1의 측정인자를 사용하여 이웃의 크기를 변경시켜가면서 비교되었다.

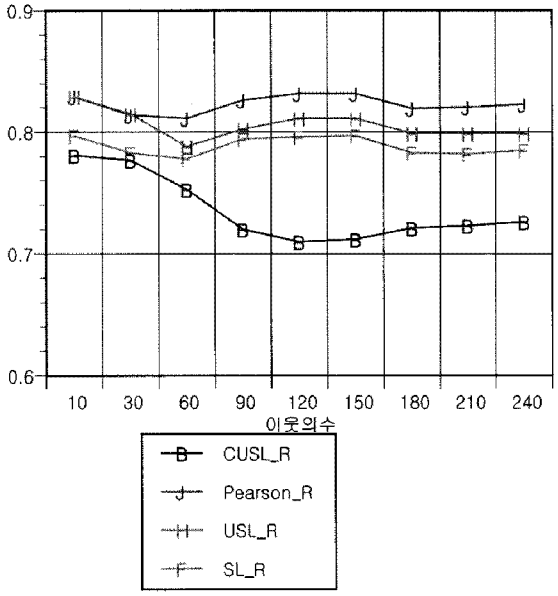

그림 3 MAE에 따른 CUSL_R, Pearson_R, USL_R, 그리고 SL_R의 성능

그림 3에서 사용자 기반과 아이템 기반을 병합한 방법인 CUSL_R은 대표적 협력적 여과 방법인 Pearson_R과 사용자 기반의 협력적 여과 방법인 USL_R의 성능보다 현저하게 높으며, 아이템 기반의 협력적 여과 방법인 SL_R보다는 다소 높다. 반면, USL_R은 사용자 기반에서 사용자를 군집하고 그룹의 특징을 추출하여 희박성의 문제를 해결한다는 장점을 갖는다.

그림 4에서 CUSL_R의 방법은 그룹의 특징을 TF-IDF를 이용하여 효과적으로 추출하고, 사용자 기반과 아이템 기반의 병합된 방법을 사용하였기 때문에 NMF를 사용하는 기존의 방법보다 정확도가 높음을 보인다. CUSL_R과 비교된 방법들은 이웃의 수가 120까지는 정확도가 점차적으로 증가하나 그 이상이 되면서 정확도가 비슷하거나 감소함을 볼 수 있다. 이와 같은 이유는 이웃의 수가 점차적으로 증가함에 따라 흥미가 비슷한

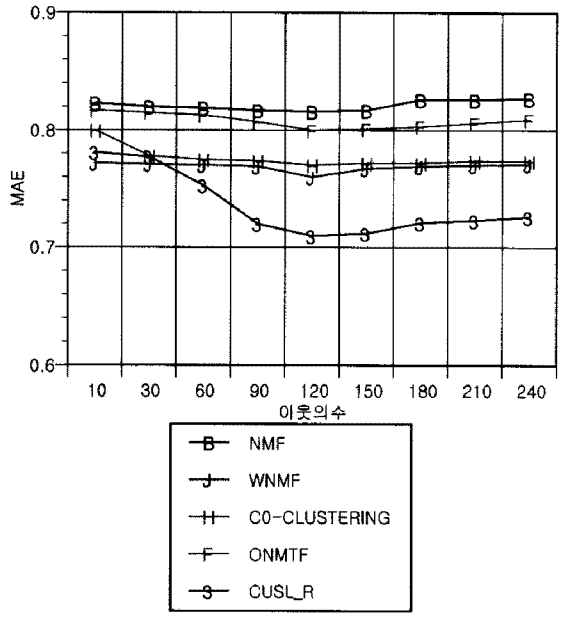

그림 4 MAE에 따른 NMF, WNMF, COCLUST, ONMTF, CUSL_R의 성능

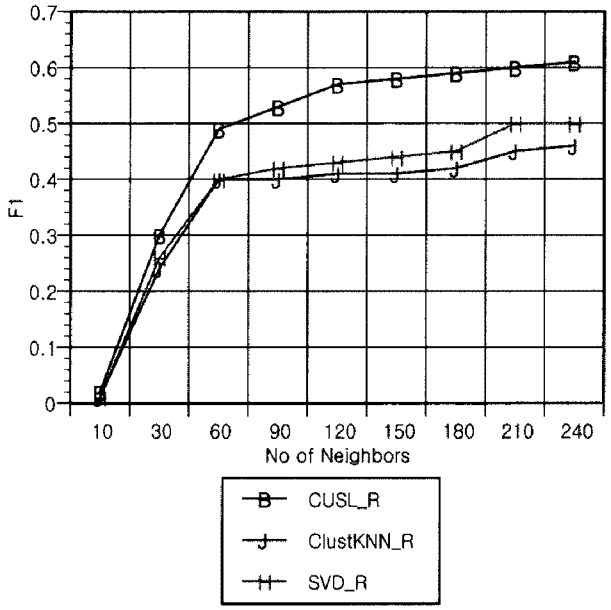

그림 5 F1 측정인자를 이용한 CUSL_R, SVD_R, 그리고 ClustKNN_R의 성능 곡선

이웃을 찾기가 용이해지고, 그 정확도도 향상되나 120이상이 되면서 이웃의 수가 너무 많아지고 유사도가 낮은 사용자가 포함되어 성능이 감소한 결과이다. 또한, 모델 기반에서 NMF만을 사용한 방법은 잡음이 그대로 포함되어 있어서 잡음을 제거하고 NMF를 사용한 방법보다 정확도가 낮으며, COCLUST 방법은 WNMF와 정확도는 비슷하나 병렬처리를 해야하는 단점을 갖는다. ONMFT의 방법은 연합 계수에 따라 정확도가 달라지며 적용한 데이터 집합에 따라 연합 계수가 다르게 적용되어야 하기 때문에 정확도면에서 낮게 나타났으며 일정하지 않다는 단점을 갖는다.

그림 5는 CUSL_R의 F1값은 SVD_R and ClustKNN_R의 방법과 비교되었다.

SVD는 사용자 선호도 값을 대상으로 학습을 하여 새로운 사용자에 대한 선호도를 예측하는 방법을 사용한다. SVD를 이용한 방법은 사용자에게 빠른 추천을 제공할 수 있으나 전체 모델을 생성하는 데 시간이 소비되며 주기적으로 모델을 갱신해야 한다는 단점이 있다. 또한, SVD를 이용한 방법은 사용자 정보를 사용하는 CUSL_R보다 정확도면에서도 낮음을 볼 수 있다. K-means 알고리즘을 사용하는 ClustKNN_R은 시간과 공간면에서 장점을 갖으나 다른 방법보다 정확도 면에서는 비효율적이다. CUSL_R은 오프라인 상에서 사용자들을 그룹으로 분류하고 온라인 상에서 사용자를 가장 정확한 그룹으로 분류한다. 분류된 그룹 안의 이웃들로부터 가장 유용한 정보를 추출해서 추천에 이용하므로 정확도가 가장 높음을 볼 수 있다.

5. 결 론

본 논문에서는 저차원 선형 모델을 사용하는 병합된 협력적 여과 방법을 제안하였다. 제안된 방법은 사용자 기반과 아이템 기반을 병합한 하이브리드 협력적 여과 기술을 사용함으로써 추천의 정확도를 향상시켰다. 또한, 사용자를 오프라인 상에서 그룹으로 군집시키고, 군집된 사용자들을 기반으로 그룹의 특징을 추출해서 온라인 상에서 새로운 사용자를 그룹으로 분류시킴으로써 사용자를 그룹으로 분류하는 데 필요한 시간을 단축시켰다. 마지막으로, NMF를 이용하여 군집된 그룹을 대상으로 TF-IDF를 적용하여 군집의 특징을 보다 효율적으로 표현할 수 있었다.

참 고 문 헌

[1] Wang, J., de Vries, A. P., and Reinders, M. J. T., "Unifying User-based and Item-based Collaborative Filtering Approaches by Similarity Fusion," In *Proceedings of SIGIR2006*, 2006.

[2] Breese, J. S., Heckerman, D., and Kadie, C., "Empirical analysis of predictive algorithms for collaborative filtering," In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 1998.

[3] Wei, Y. Z., Moreau, L., and Jennings, N. R., "Learning users' interests by quality classification in market-based recommender systems," *IEEE Trans on Knowledge and Data Engineering*, vol.17, no.12, pp.1678-1688, 2005.

[4] Sawar, B. M., Karypis, G., Konstan, J. A., and Riedl, J., "Application of dimensionality reduction

in recommender system - A case study," In *Proceedings of ACM WebKDD*, 2000.
[5] 김종훈, 김용집, 정경용, 임기욱, 이정현, "분류 속성과 Naive Bayesian을 이용한 사용자와 아이템 기반의 협력적 필터링", *한국콘텐츠학회논문지*, 제7권, 제11호, 2007.
[6] Karypis, G., "Evaluation of item-based top-N recommendation algorithms," In *Proceedings of the ACM Conference on Information and Knowledge Management*, 2000.
[7] Canny, J., "Collaborative Filtering with Privacy via Factor Analysis," In *Proceedings of the 25th ACM SIGIR*, 2002.
[8] Zhang, S., Wang, W., Ford, J., and Makedon, F., "Learning from Incomplete Rating Using Non-negative Matrix Factorization," In *Proceedings of SDM2006*, 2006.
[9] Chen, G., Wang, F., Zhang C., "Collaborative filtering using orthogonal nonnegative matrix tri-factorization," Information Processing and Management: an International Journal, vol.45, no.3, 2009.
[10] Wu, M., "Collaborative Filtering via Ensembles of Matrix Factorizations," In *Proceedings of KDD Cup and Workshop 2007*, 2007.
[11] Lee, D. and Seung, H., "Algorithms for non-negative matrix factorization," *Advances in Neural Information Processing Systems*, pp.556-562, 2001.
[12] Liu, W. and Yi, J., "Existing and New algorithms for nonnegative matrix factorization," Tech. rep., Department of Computer Sciences, University of Texas at Austin, 2003.
[13] Xu, W., Liu, X., and Gong, Y, "Document clustering based on non-negative matrix factorization," In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, 2003.
[14] George, T and Meruge, S., "A Scalable Collaborative Filtering Framework based on Co-clustering," In *Proceedings of the 5th IEEE Conference on Data Mining (ICDM)*, 2005.
[15] MovieLens collaborative filtering data set, "Http://www.cs.umn.edu/Research/GroupLens/index.html," GROUPLENS RESEARCH PROJECT, 2000.
[16] Salton, G. and McGill, M. J., Introduction to Modern Information Retrieval, McGraw-Hill, 1983.
[17] Church, K. W. and Hanks, P., "Word association norms, mutual information, and lexicography," *Computational Linguistics*, vol.16, no.1, 1990.
[18] Shannon, C. E., "A mathematical theory of communication," *Bell System Technical Journal*, vol.27, pp.379-423, 1948.
[19] Torkkola, K. and Campbell, W. M., "Mutual Information in Learning Feature Transformations," In *Proceedings of Int'l Conf. Machine Learning*, 2000.
[20] Deshpande, M. and Karypis, G., "Item-based top-n recommendation algorithms," *ACM Trans. Inf. Syst.*, vol.22, no.1, 2004.
[21] Sarwar, B., Karypis, G., Konstan, J., and Riedl, J., "Item-based collaborative filtering recommendation algorithms," In *Proceedings of the WWW Conference*, 2001.
[22] Linden, G., Smith, B. and York, J., "Amazon.com recommendations: Item-to-item collaborative filtering," *IEEE Internet Computing*, 2003.
[23] Kim, H., Lee, H., and Seo, J., "Improving FAQ Retrieval Using Query Log Clustering in semantic space," In *Proceedings of AIRS 2005*, pp.233-245, 2005.
[24] Herlocker, J., Konstan, J., Terveen, L., and Riedl, J., "Evaluating Collaborative Filtering Recommender Systems," *ACM Transactions on Information Systems*, vol.22, no.1, pp.5-53, 2004.
[25] Sarwar, B. M., Konstan, J. A., Borchers, A., Herlocker, J., Miller, B., and Riedl, J., "Using Filtering Agents to Improve Prediction Quality in the GroupLens Research Collaborative Filtering System," In *Proceedings of CSCW'98*, 1998.
[26] Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A., "Indexing by latent semantic analysis," *Journal of the american society of Information Science*, vol.41, no.6, 1990.
[27] Amershi, S. and Conati, C., "Unsupervised and supervised machine learning in user modeling for intelligent learning environments," In *Proceedings of the 2007 International Conference on Intelligent User Interfaces*, 2007.
[28] Rashid, AI M., Lan, S. K., Karypis, G., and Riedl, J., "ClustKNN: A Highly Scalable Hybrid Model& Memory Based CF Algorithm," In *Proceedings of. WebKDD*, 2006.

고 수 정
정보과학회논문지 : 소프트웨어 및 응용
제 36 권 제 4 호 참조