비부정 행렬 인수분해 차원 감소를 이용한 최근 인접 협력적 여과

고 수 정†

요 약

협력적 여과는 사용자 선호도를 예측하기 위해 그 사용자의 유형을 학습하는 데 목적을 둔 기술이다. 협력적 여과 시스템이 전자상거래에서 성공적인 기술일지라도 그들은 데이터의 고차원성과 희박성이라는 문제점을 갖는다. 본 논문에서는 이와 같은 문제점을 해결하기 위하여 비부정 행렬 인수분해(NNMF, Non-negative Matrix Factorization) 방법을 이용한 최근 인접 협력적 여과 방법을 제안한다. 행렬을 분해하기 위한 전처리로서 사용자 변동 계수를 이용하여 사용자-아이템 행렬의 결측치를 채우고, 이를 대상으로 비부정 분해 방식을 적용하여 행렬을 인수분해 한다. 비부정 분해 방식을 적용한 긍정 분해는 사용자들을 의미를 갖는 벡터로써 표현함으로써 사용자들을 의미 관계를 갖는 그룹으로 표현한다. 이와 같이 벡터로 표현된 사용자들은 백터 유사도에 의해 그들간의 유사도를 계산한다. 계산된 유사도의 정도에 의해 이웃을 결정하고, 이웃들이 평가한 아이템에 대한 흥미도를 기반으로 새로운 사용자가 평가하지 않은 아이템에 대한 결측치를 예측한다.

키워드 : 최근 인접 협력적 여과, 비부정 행렬 인수분해, 차원 감소, 사용자 변동 계수

Nearest-Neighbor Collaborative Filtering Using Dimensionality Reduction by Non-negative Matrix Factorization

Su-Jeong Ko†

ABSTRACT

Collaborative filtering is a technology that aims at learning predictive models of user preferences. Collaborative filtering systems have succeeded in Ecommerce market but they have shortcomings of high dimensionality and sparsity. In this paper we propose the nearest neighbor collaborative filtering method using non-negative matrix factorization(NNMF). We replace the missing values in the user-item matrix by using the user variance coefficient method as preprocessing for matrix decomposition and apply non-negative factorization to the matrix. The positive decomposition method using the non-negative decomposition represents users as semantic vectors and classifies the users into groups based on semantic relations. We compute the similarity between users by using vector similarity and selects the nearest neighbors based on the similarity. We predict the missing values of items that didn't rate by a new user based on the values that the nearest neighbors rated items.

Key Words : Nearest-Neighbor Collaborative Filtering, Non-Negative Matrix Factorization, Dimensionality Reduction, User Variance Coefficient

1. 서 론

협력적 여과 기술은 사용자-아이템 행렬에서 가장 비슷한 흥미를 갖는 사용자들을 찾으며, 사용자들의 이웃이 형성되었을 때 이들의 선호도를 기반으로 사용자의 선호도를 예측한다[1,15]. 협력적 여과 기술을 이용한 추천 시스템은 이들의 많은 장점으로 인하여 성공적인 평가를 받고 있으며 웹에서의 추천 시스템에서 상용적으로 사용되고 있으나 그들은 데이터 집합에서의 희박성과 고차원성이라는 단점을 갖는다[16]. 또한, 사용자-아이템 행렬에서 모든 사용자는 모든 아이템에 대하여 평가를 해야 할지라도 모든 사용자가 모든 아이템에 대하여 평가하는 경우는 없으므로 이로 인해 발생하는 결측치는 협력적 여과 시스템에서의 또 하나의 단점이다[14].

데이터 집합의 희박성과 고차원성의 단점을 해결하는 방법 중 잠재 의미 색인(Latent Semantic Indexing, LSI) 방법을 이용하는 방법이 있다[4]. 잠재 의미 색인 방법은 고차원의 벡터를 저차원 벡터로 변환시키는 방법 중 하나이다. 특

† 정 회 원 : 인덕대학 컴퓨터소프트웨어과 교수
논문접수 : 2006년 9월 25일, 심사완료 : 2006년 11월 7일

이값 분해(SVD, Singular Value Decomposition)는 잠재 의미 색인을 사용하는 대표적인 방법 중 하나이다. 특이값 분해 방법은 고차원의 벡터를 저차원의 벡터로 분해하고, 이를 기반으로 사용자들간의 의미 관계를 발견함으로써 사용자를 군집하는 데 사용하는 방법이다[11]. 특이값 분해를 사용하는 방법은 사용자-아이템 행렬의 기본 잠재 속성을 찾아내며, 대용량의 저장 공간에서도 보다 효율적으로 비슷한 사용자를 예측하여 찾음으로써 높은 성능의 추천을 제공한다는 장점을 갖는다. 그러나 이를 이용한 분해 방법은 사용자-아이템 행렬의 기본 잠재 속성에 음의 값을 포함하여 사용자들을 의미공간으로 군집시키는 것을 곤란하게 한다. 이를 해결하기 위하여 본 논문에서는 잠재 의미 색인 방법의 하나인 비부정 행렬 인수분해(Non-negative Matrix Factorization, NNMF)를 이용한다. 비부정 행렬 인수분해 방법은 대용량의 데이터에서 지수적인 함수관계를 발견하는 데 있어 매우 유용하게 사용되며, 사용자들을 군집하고자 할 때 사용자들간의 의미 관계를 발견할 수 있다는 장점을 갖는다[8,9,19]. 또한, 특이값 분해 방법과는 다르게 행렬 분해에 있어서 직교 행렬을 필요로 하지 않는다는 장점도 갖는다[8].

비부정 행렬 인수분해를 이용한 최근 인접 협력적 여과 방법은 행렬을 분해하기 위한 전처리로써 사용자-아이템 행렬의 결측치를 채운다. 행렬의 결측치를 채우기 위해 사용할 방법은 사용자 변동 계수를 이용한 방법이다[20]. 제안한 방법에서는 수식을 이용하여 자동적으로 사용자 변동 계수의 임계값을 선택하고, 그 임계값에 따라 사용자 평균에서 아이템 평균으로 전환하여 사용자들의 결측치에 대한 기본 평가값을 결정한다. 이와 같은 방법으로 사용자-아이템 행렬의 결측치를 채운 후에 그 행렬을 대상으로 비부정 분해 방식을 적용한다. 비부정 분해 방식을 적용한 긍정 분해는 사용자들을 의미를 갖는 벡터로써 표현함으로써 사용자들을 의미 관계를 갖는 그룹으로 표현한다. 이와 같이 벡터로 표현된 사용자들은 벡터 유사도에 의해 그들간의 유사도를 계산한다. 계산된 유사도의 정도에 의해 이웃을 결정하고, 이웃들이 평가한 아이템에 대한 흥미도를 기반으로 새로운 사용자가 평가하지 않은 아이템에 대한 결측치를 예측한다.

본 논문은 2장에서 사용자 변동 계수를 이용한 기본 평가값 예측 방법을 이용하여 완전 행렬을 구성하는 방법을 기술하고, 3장에서는 비부정 행렬 분해를 사용하여 사용자-아이템 행렬의 차원을 감소시키고, 이를 기반으로 사용자의 특징을 표현한다. 그리고 4장에서는 최근 인접 사용자 선택 방법을 기술하며, 5장에서는 제안된 방법과 기존의 방법의 성능을 비교하여 평가한다. 마지막으로 6장에서는 결론을 기술한다.

2. 완전 행렬 구성

비부정 분해 방식을 사용자-아이템 행렬에 적용하기 위하여 우선적으로 완전 행렬을 구성하는 전처리 작업을 한다. 이를 위하여 사용자-아이템 행렬에 있는 결측치를 기본 평가값으로 채운다[20]. 본 논문에서 사용한 방법에서는 사용자 변동 계수를 이용한 기본 평가값 예측 방법이다. 사용자 평균이나 아이템 평균 중 어느 값을 기본 평가값으로 사용할 것인가를 결정하기 위해 MovieLens 데이터 집합[10]에서 100명의 사용자씩을 임의로 선택하여 5개의 모집단을 만들었다. 이들 데이터 집합의 실제 데이터를 결측치로 가정한 후에, 이에 대한 기본 평가값으로 사용자 평균과 아이템 평균을 각각 사용하였다. 그 결과, 2개의 집단에서는 사용자 평균을 사용하였을 경우 결측치 오차가 낮아지는 결과를 보였으며, 나머지 집단에서는 아이템 평균을 사용하였을 경우 결측치 오차가 낮아지는 결과를 보였다. 이와 같은 결과는 사용자들이 아이템에 대하여 평가한 경향에 따라 두 가지 값 중 하나의 값을 기본 평가값으로 결정해야 한다는 결론을 내릴 수 있다. 변동 계수를 이용한 기본 평가값 평가의 방법에서는 사용자들이 아이템에 대하여 평가한 경향을 판단하는 기준으로 데이터 점들의 변동 정도를 나타내는 변동 계수를 사용한다. 식 (1)은 변동 계수를 계산하기 위한 식이다.

$$CV = (S / X) \times 100 \quad (1)$$

식 (1)에서 CV는 변동 계수를 의미하고, S는 표준 편차, 그리고 X는 사용자가 아이템에 대하여 평가한 값의 평균을 나타낸다.

각 사용자들에 대한 기본 평가값을 결정하기 위하여 사용자 변동 계수의 임계값을 정한다. 사용자-아이템 행렬에 속한 사용자들의 변동 계수 평균에 따라 사용자 평균에서 아이템 평균으로 전환되는 임계값이 달라지나 사용자 변동 계수의 분포 또한 그 임계값에 영향을 미친다. 따라서 임계값을 결정하기 위하여 사용자 변동 계수의 분포 정도를 나타내는 왜도(skewness of distribution)를 사용자 변동 계수 평균의 보정값으로서 사용한다. 이러한 임계값은 사용자 변동 계수의 평균과 그 분포의 규칙성을 이용하여 자동으로 결정한다. 본 논문에서는 사용자 변동 계수의 분포 정도를 구하기 위하여 식 (2)의 왜도[6]를 사용하였다.

$$skewness = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s}\right)^3 \quad (2)$$

식 (2)에서 $\bar{x}$ 는 사용자 변동 계수의 평균을 나타내고, s는 변동 계수의 표준 편차, n은 사용자-아이템에 있는 전체 사용자의 수이다.

식 (3)은 왜도가 1보다 클 경우 그 값에 로그값을 취한 결과를 나타냄으로써 임계값을 결정한다.

$$Threshold_{cv} = 1 - (CV + C\_difference)$$
$$C\_difference = \log skewness, (skewness > 0) \quad (3)$$

식 (3)에서 $Threshold_{cv}$는 사용자 변동 계수의 임계값을

나타내며, CV는 사용자 변동 계수의 평균을 나타낸다. 그리고 C_difference는 사용자 변동 계수의 보정값을 나타낸다.

반면, 왜도가 0보다 크고 1과 같거나 작은 범위에 속할 경우, 이에 대한 로그의 값은 음의 값이 나오거나 계산할 수 없는 경우가 발생한다. 이와 같은 음의 값은 평균 사용자 변동 계수에 대한 보정값으로 사용할 수 없다. 따라서 이와 같은 범위의 왜도에 대해서는 새로운 함수가 정의되어야 한다.

왜도가 0보다 크고 1과 같거나 작을 경우 식 (4)를 적용하여 평균 사용자 변동 계수에 대한 보정값을 계산한다.

$$C\_differnece = 10^{skewness - e^{0.88}} \ (0 < skewness \le 1) \qquad (4)$$

식 (3)과 식 (4)와 같이 결정한 변동 계수의 임계값을 기준으로 사용자가 평가하지 않은 결측치를 기본 평가값으로 채운다. 이를 위하여 각 사용자의 변동 계수가 임계값보다 작다면 사용자 평균을, 아닐 경우 아이템 평균을 사용한다.

<표 1>은 기본 평가값을 채우기 위한 사용자-아이템 행렬로서 10명의 사용자가 13개의 아이템에 대해 평가한 결과를 나타낸다. <표 1>에서 흑색부분은 결측치를 의미한다.

〈표 1〉 사용자-아이템 행렬의 예

	1	2	3	4	5	6	7	8	9	10	11	12	13	사용자 평균
u1		0.6	0.4	0.8	0.2	0.2	0.4	0.8	0.2	0.6	0.8	0.4	0.8	0.5167
u 2	0.8	0.4	0.4	0.8	0.8	0	1	0.6	0.6	1	0		0.6	0.5833
u3	1	0.8	0.8	0.8	0.8			0.8	0.6	0.8	0	1	1	0.7636
U4	0.2	0	0.4	0.4		0	0.6	0	0.4	0	0.4	0.4		0.2545
u5	0.6		0.4	0.8	0.4	0	0.4	1	0	0.8	0.8	1	0.4	0.5500
u6	0.8	0	0.2	1	0	0	0.6	0.4		1	1	0.4	1	0.5333
u 7	0.6	0.6	0.6	0.8	0	0.2	0.6	0.8	0.6	0.8		0.8		0.5818
u 8	1	0.8	0.8	0.8	0.2	0.4	0.8	0.8	0	0.8	1	1		0.7
u 9	1	0.8	0.8	0		0.2	0.6	0	0.8	0.8	0.8	1	0.4	0.6
u 10	0.6	0.6	1	0.8		0	0.8	0.2		0.8	0.8	0.6	0.8	0.6363
평균	0.73	0.51	0.58	0.7	0.34	0.11	0.64	0.54	0.4	0.74	0.62	0.73	0.71	0.5719

<표 2>에서 0.733, 0.764 등과 평가값은 식 (3)과 식 (4) 등을 이용한 사용자 변동 계수와 왜도를 통한 기본 평가값 예측의 결과를 나타낸다.

〈표 2〉 기본 평가값으로 채운 완전 행렬

	1	2	5	6	7	9	11	12	13	변동 계수
u1	0.733	0.6	0.2	0.2	0.4	0.2	0.8	0.4	0.8	0.4800
u2	0.8	0.4	0.8	0	1	0.6	0	0.733	0.6	0.5747
u3	1	0.8	0.8	0.764	0.76	0.6	0	1	1	0.3670
U4	0.2	0	0.343	0	0.6	0.4	0.4	0.4	0.714	0.8672
u5	0.6	0.511	0.4	0	0.4	0	0.8	1	0.4	0.6226
u6	0.8	0	0	0	0.6	0.4	1	0.4	1	0.7887
u7	0.6	0.6	0	0.2	0.6	0.6	0.622	0.8	0.714	0.4469
u8	1	0.8	0.2	0.4	0.8	0	1	1	0.714	0.46391
u9	1	0.8	0.343	0.2	0.6	0.8	0.8	1	0.4	0.6030
u10	0.6	0.6	0.343	0	0.8	0.4	0.8	0.6	0.8	0.4622

3. NNMF 차원 감소를 통한 사용자의 특징 표현

본 장에서는 2장에서와 같이 사용자-아이템 행렬을 기본 평가값으로 채운 후 NNMF를 이용하여 행렬의 차원을 감소시키고, 이를 기반으로 사용자의 특징을 의미 속성 관계로 표현한다.

3.1 NNMF 차원 감소

사용자-아이템 행렬을 R로 정의 하자. 행렬 R을 비부정 인자로 분해하기 위하여 사용자-아이템 행렬을 정규화한다. 행렬을 정규화하는 방법은 표준화 점수(z-score)를 이용하는 방법, 사용자가 아이템에 대하여 평가한 값으로부터 사용자의 평균을 빼서 그 값을 저장하는 방법이 있으며, 또한 유클리디안 길이를 이용하는 방법이 있다[14]. 평가한 값을 사용자가 평가한 모든 값의 평균으로 빼는 방법은 표준화 점수를 이용하는 방법보다 정확도가 높으나 0보다 작은 값이 나올 수 있기 때문에 비부정 인자로 분해하는 방법에는 부적당하다. 따라서 본 논문에서 제안된 방법에서는 정규화를 위하여 유클리디안 길이를 사용한다.

협력적 여과 추천에서 사용할 사용자-아이템 행렬은 $R=\{r_{ij}\}$로서 표현한다. 비부정 행렬 분해는 기저 벡터(base vector)의 집합과 은닉 벡터(hidden vector)의 집합으로부터 비부정 인자(factor)를 찾아냄으로써 행렬의 차원을 감소한다[18]. 은닉 벡터 집합의 열은 행렬 R의 열과 일대일로 대응한다. 본 논문에서는 행렬 R의 전치행렬로서 R^T를 정의한다. 비부정 분해는 $R^T \approx PC$의 근사분해이다. 행렬 P는 m_x_t 의 크기를 갖는 기저 벡터이며, 행렬 C는 t_x_n의 크기를 갖는 은닉 벡터이다. 여기서 t는 n과 m보다 작다. $P \ge 0$와 $C \ge 0$에 대해 $\|R^T - PC\|^2$의 값을 최소화한다[21]. 첫 번째 단계에서 P와 C는 임의의 비부정 행렬로 초기화된다. 다음으로 $\|R^T - PC\|^2$의 값이 수렴할 때까지 이 과정을 반복한다. 식 (5)는 행렬 P와 C를 수렴시키기 위하여 순차적으로 계산하기 위한 식이다[8].

$$P_{ia} = P_{ia}\frac{(R^T C^T)_{ia}}{(P^T C C^T)_{ia}} \qquad C_{a\mu} = C_{a\mu}\frac{(P^T R^T)_{a\mu}}{(P^T P C)_{a\mu}} \qquad (5)$$

식 (5)에 의해 R^T는 t개의 의미 속성을 가진 P와 C로 분

$$R^T = \begin{bmatrix} 0.38 & 0.466 & 0.17 \\ 0.25 & 0.414 & 0 \\ 0.16 & 0.407 & 0.31 \\ 0.56 & 0.165 & 0.35 \\ 0 & 0.305 & 0.23 \\ 0.05 & 0.176 & 0 \\ 0.19 & 0.383 & 0.48 \\ 0.56 & 0.177 & 0 \\ 0 & 0.333 & 0.33 \\ 0.48 & 0.423 & 0.09 \\ 0.55 & 0.115 & 0.31 \\ 0.31 & 0.502 & 0.24 \\ 0.36 & 0.246 & 0.57 \end{bmatrix} \begin{bmatrix} 0.59 & 0.128 & 0.12 & 0 & 0.581 & 0.51 & 0.43 & 0.461 & 0.01 & 0.24 \\ 0.14 & 0.53 & 0.63 & 0 & 0.222 & 0 & 0.32 & 0.353 & 0.74 & 0.3 \\ 0.12 & 0.176 & 0.08 & 0.93 & 0 & 0.36 & 0.12 & 0.22 & 0.03 & 0.38 \end{bmatrix}$$

(그림 1) 행렬 RT를 AP 와 AC^T로 분해

해된다. 분해가 된 행렬 R^T는 식 (6)과 같이 표현할 수 있다.

$$R^T = AP \cdot AC^T = [Ap_1,...,Ap_a,...,Ap_t][Ac_1,...,Ac_a,...,Ac_t]^T \quad (6)$$

식 (6)에서 AP는 비부정 m_x_t 차원의 행렬이며, AC는 비부정 n_x_t 행렬이다. Ap_a와 Ac_a는 각각 m과 n의 행렬요소를 갖는 a번째 열벡터이다. 열벡터 Ap_a는 $(ap_{a1},ap_{a2},\cdots,ap_{aj},\cdots,ap_{am})$의 요소를 가지며, Ac_a는 $(ac_{a1},ac_{a2},\cdots,ac_{ai},\cdots,ac_{ar})$의 요소를 갖는다. 열벡터 Ap_a와 Ac_a^T는 AP와 AC^T의 t개 의미 속성 중에서 a번째 속성에 대한 아이템과 사용자의 가중치를 나타낸다.

<표 2>를 대상으로 비부정 분해를 설명하기 위하여 t를 3으로 정의를 했다고 가정한다. t=3인 경우, <표 1>에 나타난 사용자-아이템 행렬의 결측치를 <표 2>와 같이 채운 후에, 이에 대한 전치 행렬 R^T를 분해한다. 그 결과, (그림 1)과 같이 AP 와 AC^T를 얻었다.

본 논문에서 제안한 방법에서는 아이템을 대상으로 하지 않고 사용자를 대상으로 그들의 의미 속성을 관계를 추출한다. 즉, 사용자와 아이템간의 관계가 아니고, 사용자간의 유사도를 계산하여 가장 비슷한 흥미를 갖는 이웃들을 찾는 것을 목적으로 한다. 그러나, 행렬 R^T는 t개의 의미 속성을 갖는 아이템과 사용자 모두의 관계를 표현하므로 식 (6)에서 행렬 AP의 열을 정규화하여 그들의 길이를 1로 정의함으로써 아이템이 갖는 가중치를 제거한다. 그 결과, t개의 의미 속성에 대한 아이템 가중치를 정규화를 통하여 제거함으로써 사용자들과 t개의 의미 속성만이 고려된다. 식 (7)은 식 (6)의 행렬 AP와 AC를 정규화하기 위한 식이다.

$$AP = \left[\frac{Ap_1}{\|Ap_1\|_2},...,\frac{Ap_a}{\|Ap_a\|_2},...,\frac{Ap_t}{\|Ap_t\|_2} \right]$$
$$AC = \left[Ac_1 \cdot \|Ap_1\|_2,...,Ac_a \cdot \|Ap_a\|_2,...,Ac_t \cdot \|Ap_t\|_2 \right] \quad (7)$$

AC^T는 10명의 사용자가 3개의 의미 속성에 대해 갖는 가중치를 식 (7)를 사용함에 의하여 표현할 수 있다. <표 3>은 식 (7)을 (그림 1)의 자료에 적용한 후에 얻은 결과를 보인다.

<표 3> 정규화된 배열 AC^T- 사용자의 의미 속성 가중치

	u1	u2	u3	u4	U5	u6	U7	u8	u9	u10
의미 속성1	0.752	0.165	0.148	0	0.747	0.651	0.548	0.593	0.006	0.303
의미 속성2	0.168	0.649	0.771	0.002	0.272	0.000	0.394	0.432	0.905	0.369
의미 속성3	0.131	0.186	0.082	0.98	0	0.381	0.125	0.023	0.029	0.398

3.2 사용자의 특징 표현

3.1절에서 아이템-사용자 행렬 R^T에서의 사용자와 아이템의 특징을 t개의 의미 속성에 대한 t차원 가중치 벡터로써 표현하였다. 이를 기반으로, 사용자간의 유사도는 각각의 가중치 벡터의 유사도를 계산함으로써 구할 수 있다. [19]의 연구에서는 문서의 특징을 t개의 의미 속성과의 관계로 표현할 수 있도록 비부정 행렬 분해를 이용한 문서 분류 방법을 제안하였다. 비부정 행렬 분해를 이용한 문서 분류 방법에서는 문서들을 행렬로 표현하고, 이를 비부정 행렬 인수분해를 사용하여 비부정 인자로 분해하였다. 이와 같이 행렬을 기저 벡터와 은닉 벡터 기반으로 분해하고, 벡터들로부터 문서들을 몇 몇의 가중치를 갖는 의미 속성들의 집합으로 표현한다. 이 연구에서는 문서를 의미 속성들 중에서 가장 큰 가중치를 갖는 의미 속성의 그룹으로 문서를 분류하는 방법을 제안하였다. 비부정 행렬 분해를 이용한 문서 분류 방법은 문서 분류의 정확도와 속도면에서 효율적이나 문서를 가장 큰 값의 가중치를 갖는 하나의 속성을 기반으로만 분류을 한다는 단점을 갖는다. 예를 들어, 문서를 사용자로 가정하였을 경우, <표 3>에 있는 사용자 u10은 3개의 의미 속성에 대해 0.303,0.369,0.398의 값을 갖는다. 따라서 이 방법에서 제안된 방법에 의하면 이 사용자는 의미 속성3에 가장 큰 가중치를 갖기 때문에 의미 속성3의 그룹으로 분류된다. 반면, 이 사용자는 의미 속성1이나 의미 속성2에 대하여도 거의 유사한 가중치를 갖고 있으나 다른 의미 속성이 갖는 가중치 값은 고려되지 않는다. 또한, 그룹의 수가 전체 의미 속성의 수와 같다는 또 다른 단점이 있다. 본 논문에서는 사용자를 t개의 의미 속성에 대해 정규화된 가중치를 갖는 벡터로서 사용자를 표현한다. 사용자 u_i를 t차원으로 식 (8)과 같이 표현한다.

$$u_i = (ac_{1i} \frac{\|Ap_1\|_2}{\|Ac_1\|_2},...,ac_{ai} \frac{\|Ap_a\|_2}{\|Ac_a\|_2},...,ac_{ti} \cdot \frac{\|Ap_t\|_2}{\|Ac_t\|_2}) \quad (8)$$

식 (8)과 같이 표현한 사용자를 함수로 나타내기 위하여 사용자를 함수로써 표현하면 식 (9)와 같다. 식 (9)는 u_i를 의미 속성 $(s_1,s_2,\cdots, s_a, \cdots,s_t)$에 대한 함수식으로서 표현한 결과이다.

$$u_i = F_i(s_1,s_2,...,s_t) \quad (9)$$

식 (9)에서 F_i는 사용자 u_i가 각 의미 속성에 대해 갖는

가중치로, 식 (10)과 같이 나타낸다.

$$F_i = (f_{1i}, f_{2i}, ..., f_{ai}, ..., f_{ti}) \quad (10)$$

식 (10)에서 f_{ai}는 사용자 u_i가 a번째의 의미 속성에 대해 갖는 가중치이다.

<표 3>의 사용자를 식 (8)를 이용하여 정규화한 후 식 (9)의 형태로 표현했을 때 다음과 같이 3차원 벡터로 표현할 수 있다.

u1=(0.4918,0.1079,0.1135)
u2=(0.108,0.418,0.161)
u3=(0.1,0.5,0.07)
.
.
u10=(0.198,0.237,0.344)

4. 최근 인접 사용자 선택

3장에서와 같이 사용자를 t개의 의미 속성에 대한 가중치 벡터로 표현한 후, 이를 이용하여 새로운 사용자에 대한 최근 이웃을 구할 수 있다. 사용자에 대한 최근 인접 사용자를 선택하기 위하여 벡터 공간 모델을 이용한다. 벡터 공간 모델은 정보 검색 분야에서 광범위하게 사용되고 있는 방법이다[12,13]. 상당히 많은 검색 엔진들이 웹문서의 순위를 나열하기 위해 유사도를 사용한다. 이와 같은 모델은 문서와 질어어를 벡터에 의해 표현함으로써 하나의 공간으로 사상시킨다. 내적곱과 같은 유사도 함수는 문서와 질의어 사이의 유사도를 계산하는 데 사용되어 왔다.

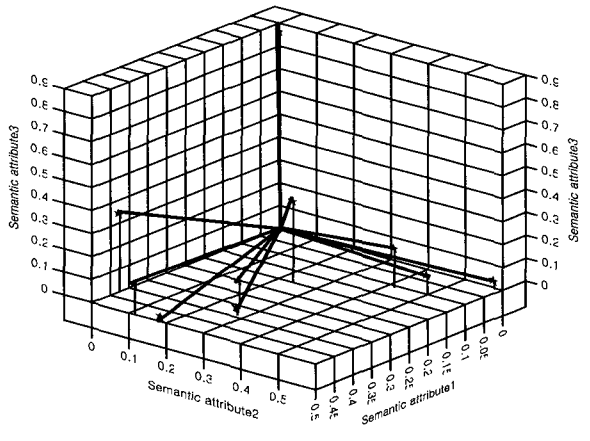
식 (11)은 식 (9)와 같이 정의된 사용자 u_i와 새로운 사용자 u_k가 t개의 의미 속성에 대한 가중치를 기반으로 이들간의 벡터 유사도를 구하는 식이다[22].

$$S(u_i, u_k) = \frac{\sum_{a=1}^{t} f_{ai} \cdot f_{ak}}{\sqrt{\sum_{a=1}^{t} f_{ai}^2} \cdot \sqrt{\sum_{a=1}^{t} f_{ak}^2}} \quad (11)$$

(그림 2)는 <표 3>의 사용자들을 벡터로 표현하여 가시화시킨 결과를 나타낸다.

식 (11)을 이용하여 기존에 있는 사용자와 새로운 사용자의 유사도를 계산하여 새로운 사용자에 대한 최근 인접의 사용자들을 찾는다. <표 4>는 <표 3>의 사용자들을 대상으로 식 (11)을 적용하여 유사도를 계산한 결과이다.

사용자 u_k에게 문서를 추천하기 위해서는 식 (11)의 $S(u_i,u_k)$의 결과 중 가장 높은 값을 나타내는 사용자 u_i의 선호도를 기반으로 아이템의 예상 선호도를 계산하고, 이 선호도가 일정 임계값 이상일 경우 추천한다. <표 1>에서와



(그림 2) 시각화를 이용한 최근 인접 여과

〈표 4〉 벡터 유사도를 이용한 사용자간의 유사도

	u2	u3	u4	u5	u6	U7	u8	u9	U10
U1	0.474	0.408	0.168	0.978	0.915	0.922	0.909	0.227	0.704
U2		0.984	0.269	0.549	0.340	0.774	0.749	0.944	0.841
U3			0.105	0.510	0.214	0.729	0.729	0.981	0.738
U4				0.001	0.505	0.183	0.039	0.034	0.642
U5					0.811	0.947	0.961	0.348	0.661
U6						0.781	0.713	0.022	0.744
U7							0.989	0.585	0.846
U8								0.595	0.763
U9									0.617

같이 사용자의 선호도는 0.2간격으로 0~1사이의 값을 나타내므로, 그 중간값인 0.5로 임계값을 설정한다. 사용자 u_k의 특정 아이템 j에 대한 선호도 p_{kj}는 식 (12)를 이용한다[1].

$$p_{kj} = \overline{r_k} + \frac{\sum_{i=1}^{n} S(u_i, u_k)(r_{ij} - \overline{r_i})}{\sum_{i=1}^{n} S(u_i, u_k)} \quad (12)$$

p_{kj}는 새로운 사용자 u_k가 문서j에 대해 예측한 선호도 값이고, $\overline{r_k}$는 새로운 사용자 u_k의 선호도 평균이다. $S(u_i,u_k)$는 새로운 사용자 u_k와 사용자 u_i의 유사도 가중치이고, n은 새로운 사용자 u_k와 다른 사용자들 간의 유사도가 0이 아닌 사용자 수이다.

<표 3>에 있는 사용자들 중 u10이 새로운 사용자라고 가정하고 새로운 아이템을 추천하고자 식 (12)에 대입할 경우의 결과는 식 (13)과 같다.

$$p_{10,5} = 3.673 \quad p_{10,9} = 0.4539 \quad (13)$$

5. 성능 평가

본 논문에서 제안한 비부정 행렬 인수분해를 이용한 최근 인접 협력적 여과 방법의 성능을 평가하기 위하여 평균 절

대 오차(MAE)[5,7]를 사용하였으며, 기존의 방법들과 매개변수에 대한 사용의 민감성을 고려하면서 그 성능을 비교하였다. 성능 평가 결과의 분석은 대응일치 t-검증(paired t-test)[3,14]과 95%의 신뢰도 수준에서 본페로니 절차(Bonferroni procedure for multiple comparison statistics)의 에이노바(ANOVA) 분석[2,17]을 사용하였다. MAE는 사용자-아이템 행렬에서 실제 평가값과 예측값 사이의 차이를 기반으로 그 정확도를 측정한다. MAE는 추천의 정확도를 측정하기 위해 가장 통상적으로 사용하는 방법이며, 가장 쉬운 방법이기도 하다. n개의 결측치를 갖는 데이터 집합에서 실제 평가 값이 a_i이고, 예측된 값이 p_i인 경우, MAE는 식 (14)에 의해 계산된다.

$$MAE = \frac{\sum_{i=1}^{n} a_i - p_i}{n} \quad (14)$$

비부정 행렬 인수분해를 이용한 최근 인접 협력적 여과 방법의 성능을 평가하기 위해 GroupLens Research Center의 MovieLens 평가 데이터를 사용하였다[10]. MovieLens 데이터 집합은 6040의 사용자들이 3960의 영화에 대해, 총 1000000의 평가를 하였다. 기존의 협력적 여과 연구는 다양한 사용자 집합을 대상으로 실험을 하였다. 예를 들어, [16]의 연구에서는 943명을 대상으로, [3]는 1400명을 대상으로, [1]에서는 5000명을 대상으로 실험을 하였다. 본 논문에서는 1000명의 사용자를 데이터 집합으로부터 무작위로 선택하였으며, 그 사용자들은 0에서 1까지 0.2의 간격으로 30개의 영화보다 더 많은 영화에 대하여 평가하였다.

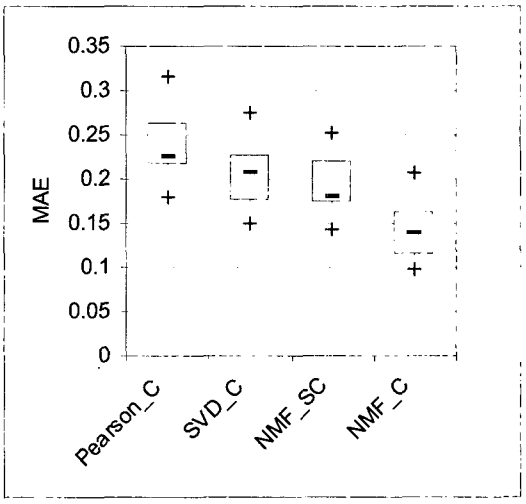
비부정 행렬 인수분해를 이용한 최근 인접 협력적 여과 방법의 성능을 평가하기 위해 10개의 데이터 집합으로 나눈다. 10개의 데이터 집합은 dataset1, dataset2,…, dataset10으로 표기하였다. 사용자의 평가 정보는 각 데이터 집합마다 다르며, 그 분포는 비부정 행렬 인수분해를 이용한 최근 인접 협력적 여과 방법의 정확도에 영향을 미친다. 따라서, 그 데이터 집합을 10개로 나눔에 의해 데이터의 분포 정보를 고려하였다. 각 데이터 집합은 훈련 데이터 집합과 테스트 데이터 집합으로 구분하였다. 훈련 데이터 집합은 사용자-아이템 행렬로부터의 실제 데이터 집합이고, 테스트 데이터 집합은 비부정 행렬 인수분해를 이용한 최근 인접 협력적 여과 방법의 정확도를 평가하기 위하여 사용되었다.

비부정 행렬 인수분해를 이용한 최근 인접 협력적 여과 방법의 성능(NMF_C)은 SVD를 이용하여 결측치를 예측하는 방법(SVD_C)[11], 비부정 행렬 인수분해를 이용하여 사용자를 속성으로 분류한 뒤에 이를 이용하여 결측치를 예측하는 방법(NMF_SC)[19] 피어슨 상관 계수를 이용하여 결측치를 예측하는 방법(Pearson_C)[15]과 비교하였다.

<표 5>는 10개의 데이터 집합에서 비부정 행렬 인수분해를 이용한 최근 인접 협력적 여과 방법의 성능을 나타낸다. 즉, SVD를 이용하여 결측치를 예측하는 방법(NMF_C), 비부정 행렬 인수분해를 이용하여 사용자를 속성으로 분류한 뒤에 이를 이용하여 결측치를 예측하는 방법(NMF_SC), 피어슨 상관 계수를 이용하여 결측치를 예측하는 방법(Pearson_C)과 비교하였을 경우 성능의 결과를 보인다.

<표 5> NMF_C, Pearson_C, SVD_C, 그리고 NMF_SC의 성능 비교

	Pearson_C	SVD_C	NMF_SC	NMF_C
dataset1	0.315364	0.275913	0.221	0.168135
dataset2	0.262964	0.206156	0.221714	0.163556
dataset3	0.210897	0.173893	0.178602	0.11381
dataset4	0.283124	0.253333	0.2514	0.206924
dataset5	0.218834	0.194455	0.172339	0.11823
dataset6	0.224639	0.214168	0.17427	0.14041
dataset7	0.216429	0.226957	0.211731	0.13479
dataset8	0.217727	0.158382	0.157589	0.1067
dataset9	0.179929	0.150461	0.143023	0.098848
dataset10	0.233889	0.223961	0.17769	0.1463
평균	0.23637943	0.207767	0.1909357	0.13977029



(그림 3) 비부정 행렬 인수분해를 이용한 최근 인접 협력적 여과 방법의 성능 비교

(그림 3)은 NMF_C, Pearson_C, SVD_C, 그리고 NMF_SC의 성능을 보이고 있으며, <표 5>를 기반으로 하였다.

(그림 3)에서 <표 5>의 각 열은 하나의 상자에 해당한다. 각 상자에서 상자 위 '+' 기호는 최대값이며, 상자 아래 '+' 기호는 최소값에 해당한다. 평균은 '-' 기호에 의해 표현되었다.

<표 5>와 (그림 3)에서 NMF_C를 이용한 방법을 사용하였을 경우 나머지 다른 방법들보다 보다 성능이 우수함을 보인다. NMF_SC를 이용한 방법은 하나의 의미 속성만으로 사용자를 표현함으로 인하여 그 정확도가 저하됨을 볼 수 있다. SVD_C를 이용한 방법은 음의 값을 갖는 잠재 속성으로 인한 오차로 인하여 정확도가 낮음을 보인다. 마지막으로, Pearson_C의 방법은 행렬의 희박성 문제로 인하여 정확도가 가장 낮음을 볼 수 있다.

<표 6>은 대응일치 t-검증과 95%의 신뢰도 수준에서의

〈표 6〉 비부정 행렬 인수분해를 이용한 최근 인접 협력적 여과 방법의 성능 비교의 신뢰도 수준

	NMF_C	NMF_SC	SVD_C	Pearson_C	(Mean)
NMF_C	----	0.04132	0.04132	0.04132	0.1398
NMF_SC	yes	----	0.04132	0.04132	0.1909
SVD_C	yes	no	----	0.04132	0.2078
Pearson_C	yes	yes	no	----	0.2364

다중 비교의 본페로니 절차의 에이노바 분석을 사용한 결과의 통계적인 신뢰성을 나타낸다. <표 6>의 오른쪽 상단부는 대응일치 t-검증을 이용한 각각의 방법들간의 신뢰도를 의미하며, 그룹들간 평균에서의 차이를 나타낸다. 왼쪽 하단부는 에이노바 분석을 통해 각 방법들간에 신뢰성이 있는가의 유무를 의미한다.

<표 6>에서 NMF_C와 NMF_SC, SVD_C, 그리고 Pearson_C의 신뢰도 유무를 나타낼 때, NMF_C는 모든 방법과의 성능 차이에서 신뢰도가 있음을 볼 수 있다. 반면, NMF_SC는 SVD_C와는 큰 차이가 없으며, Pearson_C와는 성능 차이에 있어 신뢰도가 있음을 알 수 있다. SVD_C는 Pearson_C와 성능 차이에 있어 신뢰도가 없음을 나타낸다. 결론적으로, NMF_C의 방법은 NMF_SC, SVD_C, Pearson_C와의 성능 비교에 있어서 성능이 높음을 증명할 수 있다.

6. 결 론

본 논문에서는 협력적 여과 시스템에서의 희박성을 해결하기 위하여 비부정 행렬 인수분해를 이용한 차원 감소를 통한 최근 인접 협력적 여과 방법을 기술하였다. 사용자-아이템 행렬을 대상으로 비부정 행렬 인수분해를 적용하기 위해서는 그 행렬의 기본 평가값 예측이 행해져야 한다. 이를 위하여 사용자-아이템 행렬을 대상으로 사용자 변동 계수를 이용한 기본 평가값 평가 알고리즘을 적용하여 결측치를 채웠다. 사용자 변동 계수를 이용하여 기본 평가값을 결측치를 채우는 방법은 모든 결측치를 동일한 값으로 평가하는 기존의 방법과는 다르게 사용자의 특성을 고려하였다. 즉, 사용자 변동 계수와 평가 분포의 왜도를 이용하므로 기존의 기본 평가값 평가 방법들보다 높은 정확도를 보였다. 비부정 행렬 인수분해를 이용한 최근 인접 협력적 여과 방법은 고차원의 사용자-아이템 행렬의 차원을 감소시키며, 행렬의 잠재 의미 속성 분석을 통하여 비슷한 사용자들의 그룹을 찾을 수 있었다. 성능분석 결과, SVD와 같이 차원을 감소시키면서 동시에 사용자를 군집시키는 방법과의 성능을 비교했을 때, 그 성능이 높음을 보였으며, 일반적인 협력적 여과 알고리즘과 비교하였을 경우 상당히 성능이 높음을 보였다. 또한, 비부정 행렬 인수분해를 이용하여 행렬의 차원을 축소한 뒤 무조건 하나의 의미 속성으로 사용자를 분류하는 방법보다도 높은 성능을 보였다.

참 고 문 헌

[1] John. S. Breese, D. Heckerman, and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," Proceedings of the Conference on Uncertainty in Artificial Intelligence, Madison, WI, 1998.

[2] Giuseppe Carenini, Rita Sharma, "Exploring More Realistic Evaluation Measures for Collaborative Filtering," Proceeding of AAAI 2004, 2004.

[3] Sonny Han Seng Chee, Jiawei Han, and Ke Wang, "RecTree: An Efficient Collaborative Filtering Method," Proceedings of the Third International Conference on Data Warehousing and Knowledge Discovery, September, 2001.

[4] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," Journal of the American Society for Information Science, Vol. 41, No. 6, 1990.

[5] J. Delgado and N. Ishii, "Formal Models for Learning of User Preferences, a Preliminary Report," In Proceedings of International Joint Conference on Artificial Intelligence (IJCAI-99), 1999.

[6] L. Grossi, G. Gozzi, and P. Ganugi, "Distribution Analysis of Items and Ratios in Companies?Accounts using a new iterative procedure," Proceedings of compstat2002, 2002.

[7] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl, "Evaluating collaborative filtering recommender systems," ACM Transactions on Information Systems (TOIS) archive, Vol. 22, No. 1, 2004.

[8] Lee, D. and Seung, H., "Algorithms for non-negative matrix factorization," Advances in Neural Information Processing Systems, pp. 556-562, 2001.

[9] 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.

[10] MovieLens collaborative filtering data set, Http://www.cs.umn.edu/Research/GroupLens/index.html, GROUPLENS RESEARCH PROJECT, 2000.

[11] M. H. Pryor, "The effects of singular value decomposition on collaborative filtering," Technical Report PCS-TR98338, Compute Science Department, Dartmouth College, 1998.

[12] V. Rijsbergen and C. Joost, *Information Retrieval*, Butterworths, London-second edition, 1979.

[13] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, 1983.

[14] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Application of Dimensionality Reduction in Recommender System A Case Study," Proceddings of ACM WebKDD, 2000.

[15] B. Sarwar, J. Konstan, Al Borchers, J. Herlocker, B. Miller, and J. Riedl, "Using Filtering Agents to Improve Prediction Quality in the GroupLens Research Collaborative Filtering System," Proceedings of the 1998 Conference on Computer Supported Cooperative Work, 1998.

[16] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Analysis of recommendation algorithms for e-commerce," Proceedings of the 2nd ACM conference on Electronic commerce, 2000.

[17] Rita Sharma and David Poole, "Symmetric Collaborative Filtering Using the Noisy Sensor Model," Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence, 2001.

[18] Simon Shepherd, Non-Negative Matrix Factorization, http://www.simonshepherd.supanet.com/nnmf.htm, 2004.

[19] Wei Xu , Xin Liu , and Yihong Gong, "Document clustering based on non-negative matrix factorization," Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, July 28-August 01, 2003.

[20] 고수정, "협력적 여과 시스템에서 사용자 변동 계수를 이용한 기본 평가값 예측", 정보과학회논문지, 제 32권, 11월, 2005.

[21] 김윤희, "Non-negative Matrix Factorization을 이용한 텍스트 문서 및 마이크로어레이 데이터 분석", 서울대학교 학사 학위논문, 2003.

[22] 김혜재, 손기락, "K-최근접 이웃 추천 엔진에서의 벡터 유사도 사용에 대한 실험적 분석", 정보과학회 춘계 학술발표논문집(B), 제28권, 제1호, 2001.

고 수 정

email : sjko@induk.ac.kr

1990년 인하대학교 전자계산학과 졸업(학사)

1997년 인하대학교 전자계산교육전공(석사)

2002년 인하대학교 전자계산공학과(박사)

2003년~2004년 University of Illinois at Urbana-Champaign Post Doc.

2004년~2005년 Colorado State University Research Scientist

2005년~현재 인덕대학 컴퓨터소프트웨어과 교수

관심분야 : 데이터마이닝, 정보검색, 기계학습 등