사용자-상품 행렬의 최적화와 협력적 사용자 프로파일을 이용한 그룹의 대표 선호도 추출

(Extracting Typical Group Preferences through User-Item Optimization and User Profiles in Collaborative Filtering System)

고 수 정 †

(Su-Jeong Ko)

요 약 협력적 여과 시스템은 희박성과 단지 두 고객만의 선호도에 따른 상관 관계로 추천을 제공한다는 문제점과 군집내의 가장 유사한 두 사용자만의 상관 관계에 의하여 추천을 한다는 단점이 있다. 또한, 상품의 내용을 기반으로 하지 않고 선호도만을 기반으로 하므로 추천의 정확도가 사용자에 의해 평가한 자료에만 의존한다는 문제점도 있다. 이와 같이 평가된 자료를 추천에 이용할 경우, 모든 사용자가 모든 상품에 대해 성의 있게 평가할 수는 없으므로 추천의 정확도가 낮아지는 결과를 가져온다. 따라서 본 논문에서는 엔트로피을 사용하여 사용자가 상품에 대하여 평가한 자료를 기반으로 검증되지 않은 사용자를 제외시키고, 다음으로 사용자 프로파일을 생성한 후 사용자를 군집시키며, 마지막으로 그룹의 대표 선호도를 추출하는 방법을 제안한다. 기존의 사용자 군집을 이용한 방법은 군집내의 사용자만을 대상으로 유사한 사용자를 찾으므로 희박성은 해결할 수 있으나 그 외의 단점을 해결하지 못하였다. 제안한 방법에서는 상품에 대해 평가한 선호도 뿐만 아니라 상품에 대한 정보를 반영하기 위하여 연관 단어 마이닝의 방법에 의해 협력적 사용자의 프로파일을 생성하고, 이를 기반으로 벡터 공간 모델과 K-means 알고리즘에 의해 사용자를 군집시킨다. 군집된 사용자를 대상으로 상품의 선호도와 사용자의 엔트로피를 병합함으로써 최종적으로 그룹의 대표 선호도를 추출한다. 대표 선호도를 이용한 추천 시스템은 한 사용자의 부정확한 선호도를 기반으로 추천을 하는 경우에 나타나는 추천의 부정확도 문제를 해결하며, 군집내의 가장 유사한 두 사용자만의 상관 관계에 의하여 추천을 하는 단점을 보완하고, 또한 그룹 내에 가장 유사한 사용자를 찾는 데 소요되는 시간을 절약할 수 있다는 장점을 갖는다.

키워드 : 그룹의 대표 선호도 추출, 협력적 사용자 프로파일, 엔트로피, 추천 시스템

Abstract Collaborative filtering systems have problems involving sparsity and the provision of recommendations by making correlations between only two users' preferences. These systems recommend items based only on the preferences without taking in to account the contents of the items. As a result, the accuracy of recommendations depends on the data from user-rated items. When users rate items, it can be expected that not all users can do so earnestly. This brings down the accuracy of recommendations. This paper proposes a collaborative recommendation method for extracting typical group preferences using user-item matrix optimization and user profiles in collaborative filtering systems. The method excludes unproven users by using entropy based on data from user-rated items and groups users into clusters after generating user profiles, and then extracts typical group preferences. The proposed method generates collaborative user profiles by using association word mining to reflect contents as well as preferences of items and groups users into clusters based on the profiles by using the vector space model and the K-means algorithm. To compensate for the shortcoming of providing recommendations using correlations between only two user preferences, the proposed method extracts typical preferences of groups using the entropy theory. The typical preferences are extracted by combining user entropies with item preferences. The recommender

· 이 논문은 한국과학재단의 해외 Post-doc. 연수지원에 의하여 연구되었음

† 정 회 원 : 인덕대학 컴퓨터 소프트웨어과 교수
sjko@induk.ac.kr

논문접수 : 2004년 7월 19일
심사완료 : 2005년 5월 4일

system using typical group preferences solves the problem caused by recommendations based on preferences rated incorrectly by users and reduces time for retrieving the most similar users in groups.

Key words : Extracting typical group preferences, collaborative user profiles, entropy, recommender system

1. 서 론

협력적 여과 시스템은 사용자와 가장 유사한 흥미를 나타내는 사용자의 선호도에 따라 사용자에게 필요한 정보를 제공하는 시스템이다[1]. 이러한 협력적 여과 시스템은 정보의 내용을 직접 분석할 필요없이 사용자간의 관계만을 이용하며, 정보 추천의 범위를 넓힘으로써 여러 종류의 상품을 추천할 수 있다[2]. 또한, 정보의 내용 뿐만 아니라 정보의 우수성에 따라 정보를 추천할 수도 있다는 장점을 갖는다[3].

Ringo나 GroupLens와 같은 협력적 여과 시스템은 피어슨 상관을 이용하여 사용자간의 정보에 대한 선호도를 비교하여 상관 관계를 계산하고, 이를 이용하여 유사한 사용자를 찾는다[4]. 이와 같은 시스템은 두 고객 사이의 상관 관계를 오직 그들이 모두 선호도를 표시한 상품에 대해서만 계산함으로 인하여 다음과 같은 문제점을 갖는다. 첫째, 상품의 수가 많을 경우 같은 상품에 대하여 두 고객 모두 선호도를 표시할 확률이 적으므로 추천을 제공하기 어렵다는 희박성의 단점을 갖는다[5]. 둘째, 선호도에 따른 두 고객의 상관 관계가 높지 않고, 그들의 선호도가 상대방의 선호도 예측에 좋은 자료가 될 수 있을지라도 서로의 상관 관계가 높지 않다는 이유로 두 고객의 선호도 정보는 추천에 이용되지 않는다[6]. 셋째, 상관 관계가 오직 두 고객 사이에서만 계산된다는 것이다. 두 고객과의 상관 관계만으로 추천을 한다고 할 경우, 두 고객 중 한 고객이 일부의 상품에 대해 성의없이 평가를 하였다면 나머지 고객에게는 전혀 의외의 상품이 추천될 것이다[7]. 이와 같은 여러 문제점 중 희박성을 해결하기 위하여 EM 알고리즘[8], K-means 알고리즘[9,10], 엔트로피 가중치 및 SVD를 이용한 군집의 특징 추출을 사용하여 사용자를 군집시키는 여러 연구가 있다[11-13]. 이러한 방법은 모든 사용자를 대상으로 비슷한 사용자를 찾는 것이 아니고, 상품에 대해 비슷한 유형으로 평가를 한 사용자를 같은 그룹으로 군집 시킨 후에 군집내의 사용자를 대상으로 비슷한 사용자를 찾으므로 희박성의 단점을 해결할 수는 있다[13]. 그러나 그룹내의 오직 두 고객 사이의 상관 관계에 의존하여 추천을 한다는 단점과 선호도에 따른 상관 관계가 높지 않다는 이유만으로 추천이 불가능하다는 단점을 해결하지는 못하였다[14].

본 논문에서는 추천의 정확도를 저하시키는 검증되지 않은 사용자에 의해 평가된 상품 정보를 제외시키고, 검증된 사용자들을 대상으로 프로파일을 생성한 후에 이를 이용하여 그룹의 대표 선호도를 추출하는 방법을 제안한다. 사용자가 평가한 선호도는 검증되지 않은 자료이므로 선호도의 값이 거의 동일하게 분포된 사용자는 제외시켜야 한다. 본 논문에서는 이와 같이 평가한 사용자를 '검증되지 않은 사용자'라고 정의한다. 사용자중에서 임계값 이상의 엔트로피를 갖는 사용자만을 검증된 사용자로 선정하고, 검증되지 않은 사용자가 평가한 자료는 추천에 이용되는 자료로부터 제외한다. 제안된 방법에서는 상품에 대해 평가한 선호도 뿐만 아니라 상품에 대한 정보를 반영하기 위하여 연관 단어 마이닝의 방법[15]을 사용한다. 이 방법에 의해 협력적 사용자의 프로파일을 생성[16]하고, 이를 기반으로 벡터 공간 모델과 K-means 알고리즘에 의해 사용자를 군집시킨다. 본 논문에서는 협력적 여과 시스템의 {사용자-상품} 행렬에 속한 사용자를 '협력적 사용자'의 용어로 간략하게 사용한다. 이에 따라 기존 협력적 여과 시스템의 단점인 희박성과 사용자의 선호도에 따른 상관 관계만으로 추천하는 단점을 해결한다. 또한, 군집내의 가장 유사한 두 사용자만의 상관 관계에 의하여 추천을 하는 단점을 해결하기 위하여 엔트로피[17]를 이용함으로써 그룹의 대표 선호도를 추출한다. 그룹의 대표 선호도를 추천에 이용함으로써 오직 한 사용자의 선호도를 기반으로 추천을 하는 경우에 나타나는 추천의 부정확도를 저하시키며, 그룹 내에 가장 유사한 사용자를 찾는 데 소요되는 시간을 절약할 수 있으므로 동적인 추천을 가능하게 한다.

제안된 방법은 사용자가 웹 문서에 대해서 선호도를 평가한 데이타베이스에서 평가되었으며, 기존의 방법보다 보다 효율적임을 증명한다.

2. 대표 선호도 추출을 위한 시스템 구성도

그림 1은 추천 시스템에서 협력적 사용자 그룹의 대표 선호도를 추출하기 위한 전체 구성도를 나타낸다. 그림 1은 협력적 사용자를 최적화하는 A단계, 상품으로부터 연관 단어 마이닝에 의해 특징을 추출하고, 이를 이용하여 협력적 사용자의 프로파일을 생성 시키는 B단계, 협력적 사용자를 군집시키고 각 그룹에서의 대표 선호도를 추출하는 C단계로, 총 3단계로 구성된다.

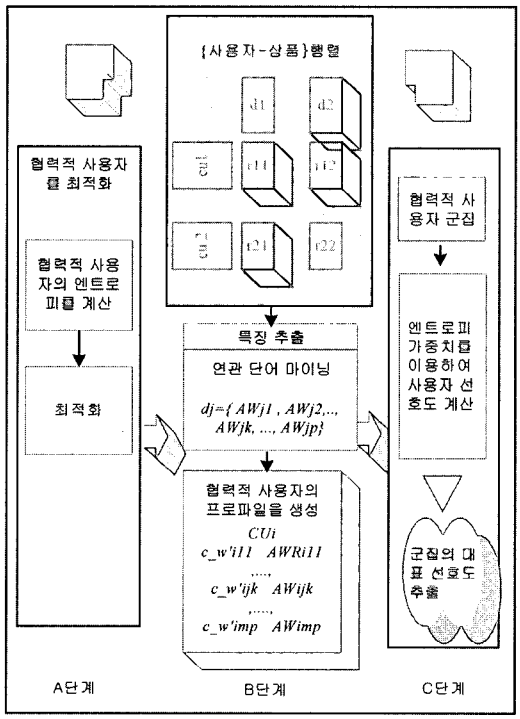

그림 1 추천 시스템에서 협력적 사용자 그룹의 대표 선호도를 추출하기 위한 전체 구성도

A단계에서는 {사용자-상품}의 행렬로부터 검증되지 않은 사용자를 제외시킴으로써 협력적 사용자를 최적화한다. 최적화시키기 위하여 그룹내의 각 사용자에 대한 엔트로피를 계산한다. 엔트로피를 계산한 후, 이를 기반으로 검증되지 않은 사용자를 제외 시킨다.

B단계에서는 {사용자-상품}의 행렬을 기반으로 협력적 사용자의 프로파일을 생성한다. 상품의 정보를 반영하기 위하여 연관 단어 마이닝을 사용함으로써 상품의 특징을 추출한다. 이에 따라 상품은 연관 단어의 집합으로 구성한다. 협력적 사용자가 상품에 대해 평가한 선호도를 연관 단어의 최초 가중치로 지정하고, 연관 단어의 빈도에 대한 통계를 계산하여 초기 가중치와 병합한다. 그 결과를 최종적으로 연관 단어의 가중치로 결정한다. 이와 같이 가중치가 부여된 연관 단어를 기반으로 협력적 사용자의 프로파일을 구성한다.

C단계에서는 가중치가 부여된 협력적 사용자 프로파일을 기반으로 사용자를 군집시키고, 군집된 각 그룹의 대표 선호도를 추출한다. 사용자를 군집시키기 위하여 우선 사용자간의 유사도를 벡터 공간 모델에 의하여 계산하고, 계산된 유사도를 대표적인 군집 알고리즘인 K-means 알고리즘에 적용함에 따라 사용자를 군집시킨다. 다음으로, 그룹의 사용자가 평가한 선호도에 해당 사용자의 엔트로피를 반영하여 선호도를 재계산한다. 마지막으로, 재계산한 선호도는 {사용자-상품}의 행렬에 적합한 값이 아니므로, 적합한 값으로 변환시키는 과정이 필요하다. 이와 같은 과정을 거쳐서 최종적으로 군집의 대표 선호도를 추출할 수 있다.

3. 협력적 사용자의 최적화

상품에 대한 사용자 평가 분포는 협력적 여과 시스템에서 상당히 중요한 의미를 차지한다. 군집내의 사용자들이 상품에 대해 평가한 대표 선호도를 추출하기 위해서는 상품에 대해 보다 성의있게 평가한 사용자의 선호도를 기반으로 하여야 한다. 본 장에서는 협력적 여과 시스템에서 {사용자-상품} 행렬의 구성 방법을 보이며, 추천의 정확도를 높이기 위하여 사용자들이 상품에 대해 평가한 데이타베이스로부터 적합하지 않은 평가를 한 사용자들을 제외시키는 방법을 기술한다.

3.1 {사용자-문서} 행렬의 구성

p개의 특징 벡터로 구성된 m개의 문서와 n명의 사용자 집합을 정의할 경우, 사용자 집합은 $U=\{cu_i\}(i=1,2,\cdots,n)$로 정의하고, 문서의 집합은 $I=\{d_j\}(j=1,2,\cdots,m)$로 정의한다. $R=\{r_{ij}\}(i=1,2,\cdots,n\ \ j=1,2,\cdots,m)$는 {사용자-문서}의 행렬이다. 행렬의 요소 r_{ij}는 문서 d_j에 대한 사용자 cu_i의 선호도를 나타낸다. 표 1은 협력적 여과에 대한 {사용자-문서}의 행렬을 보인다.

표 1 협력적 여과 시스템에서 {사용자-문서} 행렬

	d_1	d_2	d_3	d_4	⋯	d_j	⋯	d_m
cu_1	r_{11}	r_{12}	r_{13}	r_{14}	⋯	r_{1j}	⋯	r_{1m}
cu_2	r_{21}	r_{22}	r_{23}	r_{24}	⋯	r_{2j}	⋯	r_{2m}
⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
cu_i	r_{i1}	r_{i2}	r_{i3}	r_{i4}	⋯	r_{ij}	⋯	r_{im}
⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
cu_n	r_{n1}	r_{n2}	r_{n3}	r_{n4}	⋯	r_{nj}	⋯	r_{nm}

웹 문서 추천을 위한 협력적 여과 시스템에서 사용자는 문서에 대한 선호도의 정도를 평가한다. 선호도의 정도는 0~1.0까지 0.2씩 증가하면서 총 6단계로 구분한다. 6단계의 선호도 단계 중에서 0.5보다 큰 선호도로 평가를 받은 문서는 사용자가 흥미를 갖는 문서라고 평가한다.

표 1에서 r_{ij}는 식 (1)의 형태로 정의한다. 즉 행렬의 요소 r_{ij}는 선호도의 6단계와 전혀 평가를 하지 않은 경우 중 하나에 속한다.

$$F=\{0\le k\le 6 \mid f(k)\in E\}, r_{ij}\in E(i=1,2,\ldots,n)(j=1,2,\ldots,m)$$

$$E=\{\phi,0.0,0.2,0.4,0.6,0.8,1\} \quad (1)$$

식 (1)에서 ϕ는 협력적 사용자 i가 문서 j에 대해 평

가를 하지 않았음을 의미한다.

본 논문에서 사용하는 웹 문서는 웹 문서 수집기에 의해 수집된 컴퓨터에 관련된 문서이다. 웹 문서의 특징은 연관 단어 마이닝의 방법[15]을 사용하여 연관 단어 벡터 모델의 형태로 표현한다[16]. 문서의 특징을 연관 단어 벡터 모델의 형태로 표현하기 위해 Apriori 알고리즘[17]을 사용한다. Apriori 알고리즘은 형태소 분석[18]에 의해 추출된 명사들로부터 연관 단어를 마이닝한다. 마이닝한 결과를 이용하여 각 문서를 연관 단어들의 집합, 즉 연관 단어 벡터 모델로 나타낸다. 표 2는 연관 단어 마이닝에 의해 추출한 웹문서의 특징을 나타낸다.

표 2 웹문서로 부터 추출된 특징의 예

웹문서	문서의 특징
웹문서1	데이타&암호&통신망 게임&설명&제공&공략 게임&이용&기술&개발 삭제&게임&개인전&경고
웹문서2	국내&최신&기술&설치 게임&순위&이름&스포츠 게임&일정&선수&참가 위원회&선수&선발

표 3은 협력적 여과 추천 시스템에서 웹 문서에 대해 사용자가 평가한 선호도의 예이다. 표 3에서 $AWkrj$는 협력적 사용자 cu_i에 의해 선호도가 평가된 문서 d_j의 k 번째 특징이며, '?'의 의미는 선호도를 자동으로 평가해야 하는 부분임을 나타낸다.

표 3 특징으로 표현한 웹 문서에 대해 사용자가 평가한 선호도의 예

	{ AW1r1, AW2r1, ⋯, Awkr1, ⋯, Awpr1}	{ AW1r2, AW2r2, ⋯, Awkr2, ⋯, Awpr2}	⋯	{ AW1rm, AW2rm, ⋯, AWkrm, ⋯, AWprm}
cu_1	0.2	1	⋯	0.4
cu_2	?	0.8	⋯	0.6
cu_3	0.4	0.6	⋯	?
⋯	⋯	⋯	⋯	⋯
cu_n	0.4	?	⋯	?

3.2 엔트로피을 이용한 최적화

상품에 대한 사용자의 평가에 있어서 사용자가 모든 상품에 대해 모두 0.8로 평가를 하였다면, 그 사용자가 평가한 자료를 기반으로 대표 선호도를 추출한다는 것은 불가능하다. 모든 상품에 대해 모두 0.8로 평가를 하였다는 의미는 사용자가 지루하여 생각을 하지 않거나 시간이 부족한 상태에서 평가를 한 것이므로, 사용자의 선호도를 나타낸 자료라고 할 수 없다. 따라서 이를 기반으로 대표 선호도를 추출하는 것은 추천의 정확도를 저하시키는 결과를 가져온다. 즉, 사용자가 상품에 대해 매우 이산적으로 평가를 하여 상품에 대해 평가한 선호도의 값이 0~1사이에서 균등하게 나타났다면, 이 사용자가 평가한 자료는 군집 내 사용자들의 대표 선호도를 추출하는 데 매우 유용한 자료가 된다. 이와 같은 이론을 기반으로 사용자들이 상품에 대해 평가한 대표 선호도를 추출하기 위하여 엔트로피을 이용한다. 즉, 협력적 사용자 cu_i가 상품에 대해 평가한 선호도 $x1, x2, x3, \cdots$ 등이 여러 가지 값을 갖고 각각이 균일한 분포를 가질수록 불확실성이 커져서 엔트로피가 증가하는 데 반해, 확률변수가 적은 경우의 수를 갖고 한 두 가지만이 집중적으로 나타날수록 불확실성이 작아져 엔트로피는 감소한다. 표 4는 협력적 사용자 A,B,C,D,E,F,G,H가 상품 1,2,3,4,5,6에 대해서 평가한 선호도를 나타낸다.

표 4 협력적 사용자가 상품에 대해 평가한 선호도

	상품1	상품2	상품3	상품4	상품5	상품6
사용자A	0.8	0.6	0.4	0.2	1	0
사용자B	0.6	0.6	0.6	0.6	0.6	0.6
사용자C	0.2	0.4	0.2	0.4	1	0.4
사용자D	0.2	0.8	0.6	1	0.4	1
사용자E	0.6	0.6	0.8	0.6	0.6	0.8
사용자F	0.2	0.2	0.4	0.4	0.6	0.2
사용자G	1	1	1	1	1	1
사용자H	0.2	0.2	0.2	0.2	0.2	0.2

표 4의 자료를 기반으로 협력적 사용자의 엔트로피를 구한다. 식 (2)는 군집 내 사용자의 엔트로피(H_{cui})를 구하는 식이다.

$$H_{cui} = -\sum_k Rp_{cui,f(k)} \cdot \log_2 Rp_{cui,f(k)} \quad (2)$$

식 (2)에서 $Rp_{cui,f(k)}$는 협력적 사용자 cu_i가 모든 상품에 대하여 식 (1)의 집합 E의 원소 중에서 k=0인 경우를 제외하고 k번째 값인 $f(k)$로 평가될 확률을 나타낸다.

표 5는 표 4을 식 (2)에 대입한 결과로 협력적 사용자의 엔트로피를 나타낸다.

그룹의 대표 선호도 추출을 위해서는 엔트로피가 작은 협력적 사용자는 제외하고, 엔트로피가 높은 협력적 사용자만을 수집한다. 이를 위하여 협력적 사용자의 엔트로피에 대한 임계값을 정하고, 그 임계값 보다 높은 엔트로피를 갖는 사용자는 추출하고 낮은 엔트로피를 갖는 사용자는 제외한다. 200명의 사용자를 40명씩 5개의 집단으로 나누어서 임계값의 수치를 다르게 지정하여 실험해 본 결과, 집단에 속한 사용자의 분포에 따라 각기 다른 임계값에서 가장 높은 추천의 정확도를 보였

표 5 협력적 사용자의 엔트로피

	f(1) =0	f(2) =0.2	f(3) =0.4	f(4) =0.6	f(5) =0.8	f(6) =1	엔트로피
사용자A	0.166	0.167	0.167	0.167	0.167	0.167	0.778
사용자B	0	0	0	1	0	0	0.000
사용자C	0	0.333	0.5	0	0	0	0.310
사용자D	0	0.167	0.167	0.167	0.167	0.333	0.678
사용자E	0	0	0	0.667	0.333	0	0.276
사용자F	0.5	0.333	0.167	0	0	0	0.439
사용자G	0	0	0	0	0	1	0.000
사용자H	0	1	0	0	0	0	0.000
평균							0.310

다. 전반적으로, 엔트로피의 임계값을 그 집단의 평균 엔트로피보다 낮게 정할 경우, 대표 선호도 추출의 대상으로 하는 사용자가 너무 많아 선호도 평가의 분포를 이용한 대표 선호도 추출의 정확도를 저하시켰으며, 그 집단의 평균보다 큰 수치를 임계값으로 정할 경우, 대상으로 하는 사용자의 수가 너무 적어서 추출한 대표 선호도가 갖는 정확도가 저하됨을 볼 수 있었다. 따라서, 본 논문에서는 엔트로피 임계값을 대상 집단의 엔트로피 평균으로 정하여 평균보다 작은 엔트로피를 갖는 협력적 사용자는 대표 선호도 추출의 대상에서 제외하였다. 이와 같이 임계값을 집단의 평균으로 정할 경우, 표 5에서 사용자A, 사용자C, 사용자D, 사용자F가 대표 선호도 추출의 대상이 된다.

4. 사용자 프로파일의 생성과 군집의 대표 선호도 추출

협력적 여과 시스템은 사용자가 상품에 대해 평가한 자료를 이용할지라도 모든 협력적 사용자는 모든 문서에 대해 선호도를 평가하지는 않는다. 따라서 {사용자-문서} 행렬에서 결측치가 발생된다. 이러한 결측치는 {사용자-문서} 행렬을 더욱 희박하게 만드는 원인이 된다. 본 장에서는 결측치로 인한 {사용자-문서} 행렬의 희박성을 줄이고, 또한 평가 자료뿐 만 아니라 상품의 내용을 기반으로 추천을 제공할 수 있도록 3장에서 최적화 시킨 협력적 사용자를 대상으로 사용자 프로파일[16]을 생성시키고, 사용자를 군집시키는 방법을 기술한다. 이를 위해 정보검색분야에서 널리 사용되고 있는 벡터 공간 모델[19]을 이용하여 사용자간의 유사도를 구하고, 이를 기반으로 K-means 알고리즘을 사용하여 사용자를 군집시킨다. 마지막으로, 군집의 대표적인 선호도 평가를 추출하기 위해 엔트로피를 사용한다. 그룹에 속한 사용자들은 비슷한 취향을 갖는 사용자들이므로 그룹 안의 사용자들의 정보를 반영하고 있는 사용자의 엔트로피를 이용함으로써 결측이 되었거나 다소 오류가 있는 상품의 선호도를 보완할 수 있다.

4.1 가중치가 부여된 사용자 프로파일의 생성

협력적 사용자 프로파일은 {사용자-문서} 행렬을 기반으로 생성하며, 이를 위하여 문서에 대한 특징 추출이 우선되어야 한다. 문서의 특징은 [15]에서 사용한 연관 단어 마이닝을 이용한 특징 추출 방법을 사용한다.

연관 단어 마이닝을 사용함에 의해 문서 d_j는 식 (3)과 같이 연관 단어 벡터 모델로 표현된다.

$$d_j=\{ AW1rj , AW2rj,\cdots, AWkrj, \cdots, AWprj\} \quad (3)$$

협력적 사용자 cu_i의 프로파일은 식 (3)으로 정의된 각각의 특징에 가중치를 부여함으로써 생성할 수 있다. 협력적 사용자가 선호도를 낮게 평가하였을 경우, 평가된 특징에 대한 가중치는 낮게 정의되고, 선호도를 높게 평가하였을 경우 특징에 대한 가중치는 높게 정의된다. 구체적인 과정을 살펴보면 다음과 같다.

표 1에서 협력적 사용자가 문서 d_j에 대해 평가한 선호도를 r_{ij}로 정의한 것과 같이, 문서 d_j의 특징으로 추출된 각 연관 단어 $AWkrj$의 초기 가중치를 r_{ij}의 값으로 부여한다. 여기서, 연관 단어의 가중치는 $AWTkrj$로 정의한다. 식 (4)은 협력적 사용자 cu_i의 사용자 프로파일을 생성하기 위하여, 프로파일의 구성 요소인 연관 단어의 초기 가중치 $AWTkrj$를 정의하는 식이다. 식 (4)에서 연관 단어의 초기 가중치 $AWTkrj$는 사용자가 초기에 문서에 대해 평가한 선호도, {사용자-문서} 행렬의 요소로 정의한다. 사용자가 직접 평가한 선호도는 선호도를 자동으로 평가하기 위한 가장 정확하고도 중요한 자료이기 때문이다.

$$AWTkrj = Preference(AWkrj) = r_{ij}$$
$$(\text{사용자:} cu_i, 1\le k\le p,\ 1\le j\le m) \quad (4)$$

표 6은 식 (4)의 정의에 의해 생성한 연관 단어의 초기 가중치 $AWTkrj$를 계산하는 구체적인 방법을 보인다.

식 (5)와 식 (6)은 식 (4)에 따라 표 6과 같이 연관 단어 $AWkrj$에 초기의 가중치를 부여한 후, 사용자가 선호도를 평가한 모든 문서로부터 추출한 연관 단어의 빈도에 따라 가중치를 변화시키는 식이다. 식 (5)는 사용자가 선호도를 평가한 모든 문서로부터 추출한 연관 단어의 집합을 정렬하여 같은 연관 단어는 모두 $AWRj'$의 변수로써 정의하고자 하는 식이다.

$$AWRj' : AW1r1 \cdots = AWkrj \cdots = AWprm$$
$$(1\le j' \le j,\ 1\le k\le p,\ 1\le j\le m) \quad (5)$$

식 (6)은 식 (5)와 같이 정의된 연관 단어 $AWRj'$의 가중치를 정의하고 계산하는 식으로, 연관 단어 $AWRj'$의 가중치는 $AWRTj'$로 정의한다. 연관 단어에 대한 가중치 $AWRTj'$는 협력적 사용자 cu_i가 선호도를 평가한 모든 문서로부터 추출한 모든 연관 단어 집합을 검색하

표 6 프로파일 생성을 위한 초기 가중치 부여

	초기 가중치	연관 단어
d_1	$AWT1r1 = r_{i1}$	$AW1r1$
	$AWT2r1 = r_{i1}$	$AW2r1$
	...	...
	$AWTkr1 = r_{i1}$	$AWkr1$
	...	...
	$AWTpr1 = r_{i1}$	$AWpr1$
d_j	$AWT1rj = r_{ij}$	$AW1rj$
	$AWT2rj = r_{ij}$	$AW2rj$
	...	...
	$AWTkrj = r_{ij}$	$AWkrj$
	...	...
	$AWTprj = r_{ij}$	$AWprj$
d_m	$AWT1rm = r_{im}$	$AW1rm$
	$AWT2rm = r_{im}$	$AW2rm$
	...	...
	$AWTkrm = r_{im}$	$AWkrm$
	...	...
	$AWTprm = r_{im}$	$AWprm$

여 같은 연관 단어가 나타낼 때마다 가중치에 곱하도록 정의한다.

$$AWRTj' = \prod_{AWRj'=AWkrj} AWTkrj$$

(사용자 cu_i, $1 \le j' \le j$, $1 \le j \le m$) (6)

식 (6)에서 p'은 사용자 cu_i가 선호도를 평가한 문서로부터 추출한 연관 단어 집합 중에서 중복된 연관 단어를 제외한 연관 단어의 수이며, m은 사용자가 선호도를 평가한 문서의 수이다.

표 7은 식 (4)와 식 (5)의 정의에 따라 협력적 사용자 cu_i가 선호도를 평가한 연관 단어 집합에서 중복된 연관 단어의 가중치를 식 (6)에 따라 계산한 후, 연관 단어 $AWRj'$에 최종적인 가중치 $AWRTj'$가 부여되는 구체적인 방법과 예를 보인다.

식 (6)에 따라 협력적 사용자 cu_i의 프로파일 CU_i를

표 7 연관 단어에 부여된 최종 가중치

연관 단어$AWRj'$	최종가중치$AWRTj'$
$AWR1(AW1r1=AWkr1=AW1rm)$	$r_{i1} \times r_{i1} \times r_{im}$
$AWR2(AW2r1=AW2rj=AWkrm)$	$r_{i1} \times r_{ij} \times r_{im}$
$AWR3(AW3r1=AW4r3=AW5r5=AW5r7)$	$r_{i1} \times r_{i3} \times r_{i5} \times r_{i7}$
$AWR4(AWpr1=AW1rj=Awprm)$	$r_{i1} \times r_{ij} \times r_{im}$
$AWR5(AWkrj)$	r_{ij}
$AWR6(AWpr9=AW2r10)$	$r_{i9} \times r_{i10}$
...	...
$AWRj'(Awkrj=AWpr11)$	$r_{ij} \times r_{i11}$
...	...
$AWRp'(AWkrm=AWprm)$	$r_{im} \times r_{im}$

표 8 사용자 cu_i의 프로파일 CU_i의 구조($1 \le j' \le p'$)

CU_i	가중치	연관단어
1	AWRT1	$AWR1$
2	$AWRT2$	$AWR2$
...	...	...
j'	$AWRTj'_i$	$AWRj'$
...	...	...
p'	$AWRTp'$	$AWRp'$

표 8과 같이 정의한다. 표 8은 협력적 사용자 cu_i가 선호도를 평가한 모든 문서로부터 추출한 연관 단어 집합에서 중복된 연관 단어를 제외한 연관 단어 $AWRj'$에 가중치 $AWRTj'$를 부여함으로써 정의한다.

표 8에서 p'은 사용자 cu_i가 선호도를 평가한 문서로부터 추출한 연관 단어 집합 중에서 중복된 연관 단어를 제외한 연관 단어의 수이다.

4.2 벡터 공간 모델과 K-means 알고리즘을 이용한 협력적 사용자 군집

본 장에서는 사용자 프로파일을 기반으로 사용자를 군집시킨다. 벡터 공간 모델에서 저장된 모든 텍스트와 자연어 정보 요구와 같은 모든 정보는 단어의 집합, 벡터 등으로 표현되어야 한다[15]. 이와 같은 이론에서 단어는 제어 어휘로부터 추출되어야 한다. 이와 같은 제어 어휘를 구성하기 위하여 형태소 분석, 불용어 처리 등의 방법을 사용하기도 한다. 벡터 공간 모델에 따라 표 8에 의해 표현한 협력적 사용자 프로파일을 p'차원의 벡터로 정의하고, 협력적 사용자 간의 유사도 계산은 내적 유사도 함수(inner-product)를 이용한다. 내적 유사도 함수는 두 사용자간에 중첩되는 연관 단어를 검색할 수 있다. 두 협력적 사용자 프로파일에 포함되어 있는 중첩된 연관 단어가 많을수록 사용자간의 유사도는 높아지고, 적을수록 유사도는 낮아진다. 표 8에 나타난 협력적 사용자 프로파일을 기반으로 벡터 공간 모델에 의하여 사용자간의 유사도를 계산하여야 한다. 반면, 프로파일에 구성된 연관 단어의 개수가 많을 경우 구성된 개수에 의한 영향력의 불균형을 해결하기 위해 각 협력적 사용자 벡터 길이를 1로 동일하게 하는 벡터 길이 정규화 과정이 필요하다[19]. 이는 각 단어의 가중치를 가중치 제곱의 합을 그 결과의 제곱근으로 나누어 줌으로서 구할 수 있다. 식 (7)에서 n은 한 문서에서 나타난 전체 단어의 개수이다. $AWRTi$는 각 연관 단어의 가중치이며, w_i 는 조정된 각 연관 단어의 가중치이다.

$$w_1 = \frac{AWRT1}{\sqrt{\sum_{i=1}^{p'} AWRTi^2}} \qquad w_n = \frac{AWRTn}{\sqrt{\sum_{i=1}^{p'} AWRTi^2}} \quad (7)$$

이와 같은 원리에 따라 두 협력적 사용자 CU_i와 CU_j 간의 벡터 유사도를 식 (8)에 의해 구한다.

$$Sim(CU_i, CU_j) = \sum_{AWRj'} w_{j'i} \times w_{j'j} \quad (8)$$

식 (8)에서 $AWRj'$은 협력적 사용자 CU_i와 CU_j의 프로파일에 공통으로 속한 연관 단어를 의미한다. $w_{j'i}$는 협력적 사용자 CU_i의 프로파일에 포함된 연관 단어 $AWRj'$의 가중치를 의미하며, $w_{j'j}$는 협력적 사용자 CU_j의 프로파일에 포함된 연관 단어 $AWRj'$의 가중치를 의미한다. 표 9는 식 (8)에 의해 계산한 n명의 사용자 간의 유사도가 계산된 결과의 예를 나타낸다.

표 9 벡터 공간 모델에 의한 사용자 간의 유사도 계산

	CU_1	CU_2	···	CU_i	···	CU_n
CU_1	1	0.2828	···	0.567	···	0.456
CU_2	0.5721	1	···	0.826	···	0.828
···	···	···	···	···	···	···
CU_j	0.281	0.7372	···	0.4212	···	0.172
···	···	···	···	···	···	···
CU_n	0.6582	0.1285	···	0.3281	···	1

식 (8)에 의해 계산한 협력적 사용자 간의 유사도를 기반으로 K-means 알고리즘을 적용하여 사용자를 군집한다. K-means 군집 알고리즘은 데이타 분류에 있어 Maximum-Likelihood(ML) 방법의 단순화된 형태이며, 절대적 수렴에 대한 보장이 증명되지 않은 알고리즘이다. 또한, 알고리즘의 원활한 수행을 위하여 초기에 군집 해야 할 개수를 미리 정해야 하고 또 군집 중심의 초기 값에 따라 군집 된 결과의 수렴성이 달라지는 단점이 있다. 그러나 알고리즘의 간결성으로 인하여 사용자 군집에 효율적으로 응용되어 왔다[9]. K-means 알고리즘을 이용하여 사용자를 군집하는 과정은 3 단계로 구성한다. 첫번째 단계에서는 군집의 개수 K와 중심들을 초기화 한다. 두번째 단계에서는 협력적 사용자 간의 유사도를 기반으로 사용자의 소속을 구한다. 세번째 단계에서는 소속이 결정된 사용자들을 판별하기 위하여 유사도 평균의 변화치가 임계값보다 낮으면 종료한다. 본 논문에서는 200명의 사용자를 대상으로 K의 값을 2부터 100까지를 대상으로 군집을 시켜본 결과 K=8에서 가장 높은 정확도를 나타내었다. 따라서 K의 값을 8로 정하여 사용자를 군집시켰다. 또한 실험결과, 유사도 평균의 변화치는 0.01의 값으로부터 거의 일정한 값을 나타내었으므로 이 변화치를 임계값으로 결정하여 사용자를 군집시켰다. 만일 그렇지 않다면 평균과 가장 근사한 사용자를 중심으로 2단계로 가서 반복 한다.

4.3 군집 대표 선호도 추출

상품에 대한 그룹의 대표 선호도를 추출하기 위해 사용자가 상품에 대해 평가한 선호도에 사용자의 엔트로피 가중치를 곱한다. 같은 군집으로 분류된 그룹 내 사용자들은 다른 군집의 사용자들과 비교할 때 비슷한 흥미를 갖고 있으므로, 그룹 내 다른 사용자들의 정보를 이용할 경우 상품에 대해 평가된 정보를 보완할 수 있다. 사용자가 상품에 대해 평가한 정보는 엔트로피를 이용하여 추출한다. 이와 같은 사용자의 엔트로피를 상품의 선호도에 적용함으로써 오류가 있는 상품의 선호도를 보완할 수 있다. 이를 위한 식은 식 (9)와 같다. 식 (9)는 상품 d_j의 대표 선호도(Rd_j)를 추출한다.

$$Rd_j = \sum_i p_{cui,j} \cdot H'_{cui} \quad (9)$$

식 (9)는 군집내의 모든 협력적 사용자가 상품 d_j 에 대해 평가한 선호도에 유클리디언 길이[19]를 이용하여 정규화한 협력적 사용자의 엔트로피(H'_{cui})를 곱하여 모두 더한 값이다.

표 10은 협력적 사용자가 평가한 선호도, 정규화된 사용자의 엔트로피, 그리고 이를 기반으로 식 (10)에 의하여 계산된 상품별 대표 선호도를 나타낸다.

표 10 사용자의 엔트로피 가중치 및 상품의 대표 선호도

	상품1	상품2	상품3	상품4	상품5	상품6	엔트로피	정규화 엔트로피 (H'_{cui})
사용자A	0.8	0.6	0.4	0.2	1	0	0.778	0.447
사용자C	0.2	0.4	0.2	0.4	1	0.4	0.310	0.071
사용자D	0.2	0.8	0.6	1	0.4	1	0.678	0.339
사용자F	0.2	0.2	0.4	0.4	0.6	0.2	0.439	0.142
대표 선호도	0.46	0.59	0.45	0.51	0.73	0.39		

마지막으로, 식 (10)에 의해 계산된 대표 선호도 값을 추천 가능한 대표 선호도로 변환하는 작업을 행한다. 표 9에 나타난 대표 선호도는 식 (1)에서 정의된 값인 0,0.2,0.4,0.6,0.8,1의 값이 아니므로 엔트로피 가중치를 적용한 대표 선호도를 추천에서 사용할 대표 선호도로 변환하는 작업이 필요하다. 추천에서 사용할 상품d_j의 대표 선호도(Rd_j')를 구하는 알고리즘은 그림 2와 같다. 그림 2에서 $INT(\ )$함수는 소수점이하의 값을 버리고 정수만을 취하는 함수이며, $REMAINDER(A,B)$는 A를 B로 나눈 결과의 나머지 값을 취하는 함수이다.

표 11은 표 10에 나타난 상품의 대표 선호도를 그림 2의 알고리즘을 적용하여 추천에 사용하는 대표 선호도로 변환한 결과를 나타낸다.

표 11 추천에 사용할 대표 선호도로의 전환

	상품1	상품2	상품3	상품4	상품5	상품6
대표 선호도	0.4	0.6	0.4	0.6	0.8	0.4

```
temp= INT( Rdj x 10 )
temp1= REMAINDER(temp,2)
If (temp1 ==1)
 Rdj'=(temp+1)/10
 Else
Rdj'=temp/10
 Endif
```

그림 2 추천에 사용할 상품 d_j의 대표 선호도(Rd_j') 계산 알고리즘

5. 성능 평가

협력적 여과 추천을 위한 데이타베이스는 200명의 사용자와 1600개의 서로 다른 웹 문서로 구성한다. 사용자는 1600개의 웹 문서에 대해 적어도 10개의 평가를 한 사용자들이다. 1600개의 웹 문서는 웹 문서 수집기에 의해서 컴퓨터 분야의 URL로부터 수집된 문서이다. 1600개의 훈련 문서는 수작업으로 8개의 컴퓨터 분야로 분류한다. 여기서 8개의 클래스는 {게임, 그래픽, 뉴스와 미디어, 반도체, 보안, 인터넷, 전자출판, 하드웨어}의 레이블이다. 8개의 클래스로 분류한 기준은 알타비스타, 야후 등의 기존의 검색 엔진이 컴퓨터 분야의 주제를 대상으로 분류한 통계에 따른 것이다. 실험 문서는 한국어의 기본 문법 체제를 기반으로 하여야 하며, 내용의 구성이 한국어 문장으로 구성되어 있어야 한다. 또한 몇몇 단어만으로 구성되어 거의 빈문서와 같은 문서는 오류 문서이므로 실험의 대상에서 제외되어야 한다. 200명의 사용자 중 100명의 사용자는 훈련을 위해 사용하며, 나머지 100명은 테스트를 위해 사용한다.

추천의 성능을 평가하기 위해 협력적 여과 시스템의 성능 평가[20]로 많이 사용되는 MAE(Mean Absolute Error)와 순위 스코어 측정(Rank scoring metric)을 사용한다. MAE는 단일 문서의 추천 시스템을 평가하는 데 사용하며, 순위 스코어 측정은 순위가 있는 문서의 목록을 추천하는 시스템의 성능을 평가하는 데 사용한다. MAE에서 예측의 정확도는 실제로 사용자가 평가한 값과 예측된 값의 차이에 대한 절대값의 평균을 나타내며 식 (10)에 의해 정의된다.

$$s_a = \frac{1}{m_a}\sum_{j\in p_a} | p_{a,j} - v_{a,j} | \qquad (10)$$

식 (10)에서 $p_{a,j}$는 예측된 선호도이며 $v_{a,j}$는 실제로 사용자가 평가한 선호도이다. 또한 m_a는 새로운 사용자에 의해 평가된 문서의 수를 의미한다.

순위 스코어 측정은 순위가 있는 목록에 있는 문서를 사용자가 방문 또는 평가하는가의 측정이다. 순위 스코어 측정은 문서를 선택할 확률이 목록의 하단으로 갈수록 지수적으로 감소한다는 전제에서 측정된다. 각 문서는 사용자 선호도의 가중치의 값에 따라 내림차순으로 j에 의해 정렬되어 있다고 가정한다. 식 (11)은 순위가 부여된 문서의 목록에 대한 사용자 U_a의 순위 스코어 측정에 대한 기대 이용도(Expected utility)를 계산하기 위한 식이다.

$$R_a = \sum_j \frac{\max(V_{a,j} - d, 0)}{2^{(j-1)/(\alpha-1)}} \qquad (11)$$

식 (11)에서 d는 문서에 대한 중간 평가값이며, a는 반감기(halflife)이다. 반감기는 사용자가 평가하거나 방문할 50-50의 기회가 있는 목록에 있는 문서의 수이다. 본 논문의 평가에서는 반감기를 5로 사용한다. 식 (12)는 순위 스코어 척도를 사용하여 새로운 사용자에 대한 예측의 정확도를 나타내는 식이다.

$$R = 100 \times \frac{\sum_u R_u}{\sum_u R_u^{\max}} \qquad (12)$$

식 (12)에서 $R_u^{\max}$는 사용자가 평가하거나 방문한 문서가 순위가 있는 목록상에서 상위에 나타났을 경우에 측정된 순위 스코어 측정에 대한 기대 이용도의 최대값이다.

본 논문에서는 평가를 위해 제안된 대표 선호도 추출을 이용한 추천 방법(R_P_R), K-means 사용자 군집을 이용한 추천 방법(K_C_R)[10], 군집 특징 선택을 이용한 추천 방법(E_S_R)[13] 등의 방법들을 군집된 사용자의 수를 변화시키면서 성능을 비교하였다. 또한, 제안된 대표 선호도를 이용한 추천 방법(R_P_R)과 선호도 예측을 위해 피어슨 상관계수를 이용한 기존의 메모리 기반 방법(P_M)[4]을 사용자가 문서에 대해 평가한 횟수를 변화시켜가면서 비교하였다.

표 12는 식 (11)와 식 (12)를 기반으로 군집된 사용자의 수를 변화시킴에 따른 본 논문에서 제안한 R_P_R, 기존의 방법인 K_C_R과 E_S_R의 MAE와 순위 스코어를 나타낸다.

표 12 사용자 수의 변화에 따른 MAE와 순위 스코어

사용자수	MAE			Rank scoring		
	K_C_R	E_S_R	R_P_R	K_C_R	E_S_R	R_P_R
10	0.235	0.240	0.247	63.9	63.1	62.2
20	0.234	0.235	0.235	64.1	63.3	63
30	0.229	0.232	0.229	64.1	64.2	64
40	0.225	0.222	0.216	64.5	65.2	65.2
50	0.217	0.199	0.192	64.9	65.3	65.8
60	0.195	0.184	0.178	65.1	66.1	66.5
70	0.189	0.179	0.173	66.1	66.8	67.1
80	0.180	0.170	0.164	67.3	68.1	69.1
90	0.174	0.157	0.151	67.8	69.1	70.3
100	0.170	0.144	0.134	68	70.9	73.1

그림 3과 그림 4는 표 12를 기반으로 한 사용자의 수에 따른 MAE와 순위 측정 척도를 나타낸다. 그림 3과 그림 4는 사용자들의 수가 적을 경우 R_P_R의 추천 정확도는 K_C_R이나 E_S_R과 비교할 경우 다소 낮으나, 사용자의 수가 많아짐에 따라 R_P_R의 정확도는 점차 높아짐을 보인다. K_C_R의 경우, 사용자의 수가 적을 경우 다른 방법에 비해 높으나 사용자의 수가 점차 증가하면서 현저히 정확도가 낮아짐을 볼 수 있다. 이와 같은 결과로 보면, R_P_R의 경우 초기 평가 문제에서 다소 단점을 나타낸다. 따라서 초기 평가 문제를 해결함으로써 사용자들의 수가 적을 지라도 추천의 정확도가 높은 연구가 필요하다.

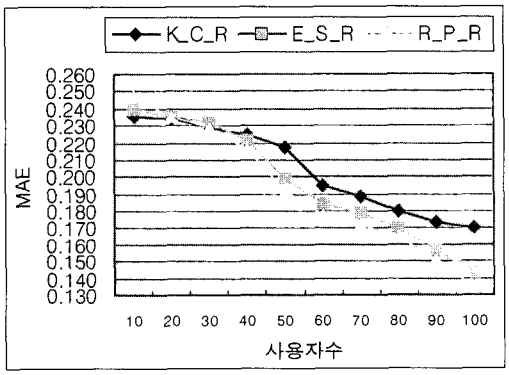

그림 3 사용자 수의 변화에 따른 MAE

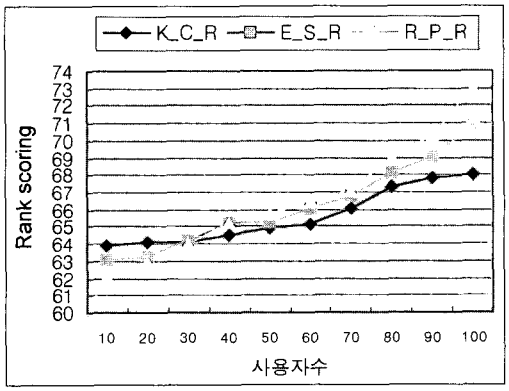

그림 4 사용자 수의 변화에 따른 순위 측정

그림 5는 사용자 수를 변화시켰을 경우 추천에 소요되는 시간을 나타내는 그림이다.

사용자의 수가 커짐에 따라 K_C_R이 추천을 위해 소요되는 시간이 증가하나 R_P_R의 경우, 사용자의 수가 늘어날지라도 소요되는 시간에 크게 소요되지 않음을 보인다. 따라서 사용자수가 작을 경우 다른 방법에 비하여 추천의 정확도가 낮은 문제는 추천에 소요되는 시간을 단축시킴으로써 단점을 보완한다.

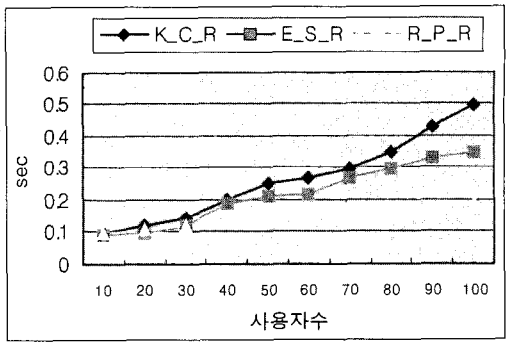

그림 5 추천에 소요되는 시간

표 13은 식 (10)과 식 (12)를 기반으로 문서에 대해 평가한 횟수를 증가시킴에 따른 제안된 방법 (R_P_R)과 P_M방법의 MAE와 순위 스코어를 나타낸다.

표 13 n번째 평가 횟수에 따른 MAE와 순위 스코어

n번째 평가	MAE		Rank scoring	
	P_M	R_P_R	P_M	R_P_R
10	0.221	0.231	63.2	63.2
20	0.217	0.226	63.9	64.0
30	0.212	0.219	64.3	64.2
40	0.210	0.207	64.9	64.8
50	0.198	0.198	65.6	65.3
60	0.190	0.185	65.9	65.8
70	0.187	0.176	66.1	66.5
80	0.186	0.165	66.9	67.9
90	0.186	0.16	67.1	69.1
100	0.185	0.157	67.81	72.3

그림 6과 그림 7은 표 13을 기반으로 한 사용자가 평가한 횟수를 증가시킴에 따른 MAE와 추천에 소요되는 시간을 나타낸다.

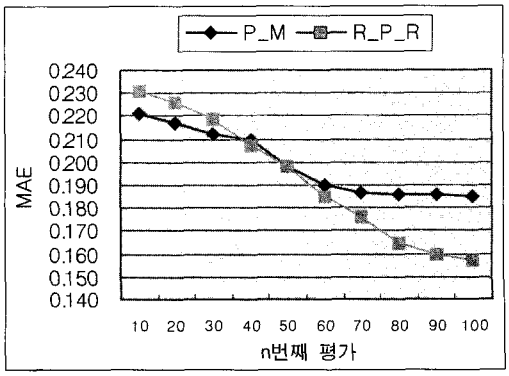

그림 6 n번째 평가에서의 MAE

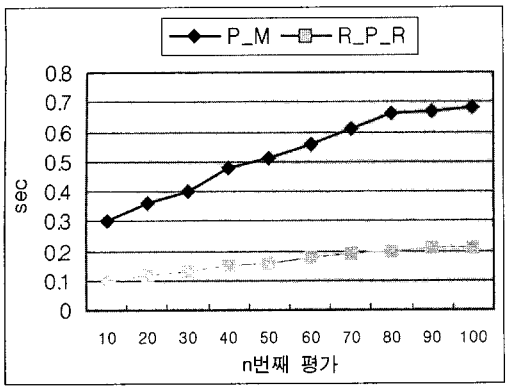

그림 7 n번째 평가에 따른 소요시간

그림 7은 평가의 수가 적을 경우 P_M의 정확도가 높으나 평가의 수가 증가함에 따라 R_P_R의 정확도가 높아짐을 보인다. 그림 7은 R_P_R의 방법이 P_M의 방법보다 추천에 소요되는 시간이 상당히 절약됨을 보인다. 그림 7 에서 보이는 R_P_R의 성능은 평가의 수가 작을 경우 나타나는 추천의 부정확도 문제는 있으나 시간에 소요되는 시간의 단축으로 인하여 전반적으로 P_M의 성능보다 높음을 보인다. 그러나 평가의 수가 작을 경우 P_M의 방법보다 정확도가 낮은 문제는 보다 많은 연구를 필요로 하는 문제이다.

6. 결 론

협력적 여과 기술은 사용자의 선호에 맞추어 사용자에게 상품의 추천을 제공하는 효율적인 기술이나 그들이 갖는 여러 문제점으로 인하여 추천의 정확도가 낮아진다는 단점을 갖는다. 본 논문에서는 이러한 문제점을 해결하기 위하여 {사용자-상품} 행렬의 최적화와 협력적 사용자 프로파일을 이용한 그룹의 대표 선호도 추출 방법을 제안하고 기존의 방법들과의 성능을 비교하였다. 제안된 방법은 기존의 방법들과 비교하여 다음과 같은 장점을 갖는다. 첫째, {사용자-상품}의 행렬로부터 상품의 선호도가 거의 동일하게 분포된 사용자, 즉 검증되지 않은 사용자를 추천의 대상으로부터 제외시켰다. 둘째, 연관 단어 마이닝의 방법에 의해 협력적 사용자의 프로파일을 생성함으로써 상품의 내용을 추가하여 사용자의 선호도에 따른 상관 관계만으로 추천하는 단점을 해결하였다. 셋째, 이를 기반으로 벡터 공간 모델과 K-means 알고리즘에 의해 사용자를 군집시킴으로써 {사용자-상품} 행렬의 희박성 문제로 인해 발생하는 추천의 오류를 저하시켰다. 넷째 군집된 사용자의 엔트로피를 계산하고, 이를 이용하여 그룹의 대표 선호도를 추출함으로써 두 사용자만의 상관 관계에 의하여 추천을 한다는 단점을 해결하였다. 다섯째, 그룹 내에 가장 유사한 사용자를 찾는 데 소요되는 시간을 절약함으로써 동적인 추천을 가능하게 하였다.

참 고 문 헌

[1] W. S. Lee, "Collaborative learning for recommender systems," In Proceedings of the Conference on Machine Learning, 1997.

[2] 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), Stockholm, Sweden, July, 1999.

[3] Badrul Sarwar, George Karypis, Josephp Konstan, and John Ridedl, "Analysis of Recommendation Algorithms for E-Commerce," Proc. Of The ACM E-Commerce 2000, 2000.

[4] A. Kohrs and B. Merialdo, "USING CATEGORY-BASED COLLABORATIVE FILTERING IN THE ACTIVE WEBMUSEUM," Proceedings of the IEEE International Conference on Multimedia and Expo-Vol. 1, 2000.

[5] L. H. Ungar and D. P. Foster, "Clustering Methods for Collaborative Filtering," AAAI Workshop on Recommendation Systems, 1998.

[6] Sarwar, B. M., Karypis, G., Konstan, J. A., and Riedl, J., "Application of Dimensionality Reduction in Recommender System-A Case Study," In ACM WebKDD 200 Web Mining for E-Commerce Workshop, 2000.

[7] C. Basu, H. Hirsh, and W. W. Cohen, "Recommendation as classification:Using social and content-based information in recommendation," In proceedings of the Fifteenth National Conference on Artificial Intelligence, pp. 714-720, Madison, WI, 1998.

[8] G. J. McLachlan and T. Krishnan, The EM Algorithm and Extensions, New York: John Wiley and Sons, 1997.

[9] K. Alsabti, S. Ranka, and V. Singh, "An Efficient K-Means Clustering Algorithm," http://www.cise.ufl.edu/~ranka/, 1997.

[10] 박지선, 김택헌, 류영석, 양성봉, "추천 시스템을 위한 2-way 협동적 필터링 방법을 이용한 예측 알고리즘", 한국정보과학회, Vol. 29, No. 9, pp. 669-675, 2002.

[11] I. Soboroff and C. Nicholas, "Combining content and collaboration in text filtering," In Proceedings of the IJCAI'99 Workshop on Machine Learning in Information filtering, pp. 86-91, 1999.

[12] D. Billsus and M. J. Pazzani, "Learning collaborative information filters," In proceedings of the International Conference on Machine Learning, 1998.

[13] 이영석, 이수원, "엔트로피 가중치 및 SVD를 이용한 군집 특징 선택", 정보과학회 논문지:소프트웨어 및 응용, 제29권, 제4호, 2002.

[14] M. Pazzani, D. Billsus, Learning and Revising User Profiles: The Identification of Interesting Web Sites, Machine Learning, Kluwer Academic Publishers, pp. 313-331, 1997.
[15] S. J. Ko and J. H. Lee, "Feature Selection using Association Word Mining for Classification," In Proceedings of the Conference on DEXA2001, LNCS2113, pp. 211-220, 2001.
[16] 고수정, 최성용, 임기욱, 이정현, "내용 기반 협력적 여과 시스템에서 사용자 프로파일을 이용한 자동 선호도 평가", 정보과학회 논문지, 제31권, 제8호, 2004.
[17] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," Proceedings of the 20th VLDB Conference, Santiago, Chile, 1994.
[18] 인하대학교, 사용자 중심의 지능형 정보 검색 시스템, 최종 연구 개발 보고서, 정보통신부, 1997.
[19] V. Rijsbergen and C. Joost, Information Retrieval, Butterworths, London-second edition, 1979.
[20] John. S. Breese and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," Proceedings of the Conference on Uncertainty in Artificial Intelligence, Madison, WI, 1998.

고 수 정
1990년 인하대학교 전자계산학과 졸업(학사). 1997년 인하대학교 전자계산교육 전공(석사). 2002년 인하대학교 전자계산공학과(박사). 2003년~2004년 University of Illinois at Urbana-Champaign Post Doc. 2004년~2005년 Colorado State University Research Scientist. 2005년 3월~현재 인덕대학 전임강사. 관심분야는 데이타마이닝, 정보검색, 기계학습