내용 기반 여과와 협력적 여과의 병합을 통한 추천 시스템에서 조화 평균 가중치

(Harmonic Mean Weight by Combining Content Based Filtering and Collaborative Filtering in a Recommender System)

정 경 용 † 류 중 경 †† 강 운 구 ††† 이 정 현 ††††

(Kyung-Yong Jung) (Joong-Kyung Ryu) (Kang-Un Gu) (Jung-Hyun Lee)

요 약 전자 상거래 분야에서 증가하고 있는 정보들 중에 사용자가 자신의 기호에 맞는 정보들 만을 선택하기 위해서 각 정보를 일일이 검토하기 어려운 일이다. 이를 보완하기 위해 정보 여과 기술이 사용되는데 최근 추천 시스템은 협력적 여과 시스템의 희박성과 초기 평가 문제를 해결하기 위해서 내용 기반 여과 시스템과 협력적 여과 시스템을 병합하는 방법을 사용한다. 본 논문에서는 혼합형 추천 시스템에서의 예측의 정확도를 향상시키기 위해서 조화 평균 가중치(CBCF_harmonic_mean)를 사용자 유사도 가중치를 구할 때 사용한다. 내용 기반의 성능을 고려하여 임계치 값을 45로 설정한 후, n/45의 Significance weight을 사용자 유사도 가중치에 적용한다. 제안된 방법의 성능을 평가하기 위해서 기존의 협력적 여과 시스템과 내용 기반 여과 시스템을 병합한 방법과 비교 평가 하였다. 그 결과 기존의 협력적 여과 시스템의 문제점을 해결하여 예측의 정확도를 높이는데 효과적임을 확인하였다.

키워드 : 협력적 여과, 내용기반 여과, CRM, 네이브 베이지안, 추천 시스템

Abstract Recent recommender system uses a method of combining collaborative filtering system and content based filtering system in order to slove the problem of the Sparsity and First-Rater in collaborative filtering system. In this paper, to make up for the prediction accuracy in hybrid Recommender system, the harmonic mean weight(CBCF_harmonic_mean) is used for calculating the user similarity weight. After setting up the threshold as 45 considering the performance of content based filtering, we apply significance weight of n/45 to user similarity weight. To estimate the performance of the proposed method, it is compared with that of combing both the existing collaborative filtering system and the content-based filtering system. As a result, it confirms that the suggested method is efficient at improving the prediction accuracy as solving problems of the exiting collaborative filtering system.

Key words : Collaborative Filtering, Content Based Filtering, CRM, Naive Bayes, Recommender System

1. 서 론

전자 상거래 분야에서 정보 기술이 발전함에 따라 사용자가 접할 수 있는 정보의 양은 기하 급수적으로 늘어났다. 특히 최근에는 인터넷의 발달로 인하여 보다 다양하고 폭 넓은 정보들이 디지털 형태로 빠르게 생산, 배포 되고 있다. 사용자가 이러한 정보 과잉 속에서 자신이 원하는 정보를 효율적으로 이용할 수 있도록 제어하고 여과하는 일을 도와주는 정보 여과 시스템이 등장하였으며, 더 나아가 사용자가 원하는 정보를 예측하고 추천함으로써 정보 과잉의 문제를 해결 하려는 노력이

† 비 회 원 : 인하대학교 전자계산공학과
kyjung@gcgc.ac.kr
†† 비 회 원 : 대림대학 컴퓨터정보과 교수
jkkyu@daelim.ac.kr
††† 비 회 원 : 가천길대학 뉴미디어과 교수
ugkang@gcgc.ac.kr
†††† 종신회원 : 인하대학교 컴퓨터공학부 교수
jhlee@inha.ac.kr
논문접수 : 2002년 8월 21일
심사완료 : 2002년 12월 28일

진행되고 있다. 이를 보완하기 위해 자동화된 정보 여과 기술이 사용되는데 대표적인 방법들로 내용 기반 여과(Content Based Filtering)와 협력적 여과(Collaborative Filtering)가 있다.

대부분의 협력적 여과 시스템들은 아이템의 수가 많아질수록 사용자가 아이템에 관련된 정보를 얻는데 어느 정도 한계가 있기 때문에 같은 아이템에 대해서 두 사용자간에 선호도를 표시할 확률은 적어지게 되고, 상관관계를 비교할 아이템의 수는 증가하게 된다. 이러한 협력적 여과 시스템은 세 가지 단점을 갖는다[1,2,3,4,5]. 첫째 사용자가 평가를 하지 않은 아이템들은 사용자에게 추천되지 않는다는 초기 평가 문제(First-Rater Problem)이다. 둘째, 대부분의 사용자들은 모든 상품에 대해 평가하지 않기 때문에 사용자-아이템의 데이타 집합은 희박한 특성(Sparsity Problem)을 보인다는 것이다. 셋째, 아이템의 속성에 대한 사용자의 선호도를 직접적으로 반영하지 못하는 문제점도 있다. [6,7,8,9,10,11]은 이러한 협력적 여과 방법의 단점을 해결하기 위해서 내용 기반 여과와 협력적 여과를 병합한 방법을 사용하였다. [5,7,8,10]의 방법들은 초기 평가 문제를 해결하였으나 희박성 문제를 해결하지 못하였다.

LSI[11]는 입력 데이타의 차원을 축소하여 축소된 차원에서 협력적 여과에 응용을 제안하였으며, 아이템의 개수 또는 사용자의 수가 많을 때 이점을 나타내고 있다. SVD[9]는 다차원의 아이템을 분류함으로써 축소된 아이템 행렬 차원에서 협력적 여과에 응용을 제안하였으며, 사용자에게 흥미로운 하나의 아이템에 관련 상품까지 추천을 한다. LSI, SVD 분류를 사용한 방법은 데이타 차원의 수를 줄임으로써 협력적 여과 시스템의 희박성 문제를 해결하였으나 초기 평가 문제는 해결하지 못하였다. 반면, Pazzani[6]가 제안한 방법에서는 새로운 사용자에 대한 예측을 사용자간의 개인화 에이전트를 협력적 여과에 응용함으로써 초기 평가 문제와 희박성 문제를 동시에 해결하려는 시도를 하였다.

2. 내용 기반 여과와 협력적 여과 시스템

2.1 내용 기반 여과 시스템

내용 기반 여과를 하기 위해서 우선적으로 텍스트 범주화 문제를 다루어야 한다. 텍스트 범주화 문제를 처리하기 위해 학습 문서에서 학습한 결과를 이용하여 문서에 적합한 범주를 할당한다. 본 논문에서는 사용자가 영화를 평가한 테이블에서 사용자의 프로파일을 학습하기 위해서 Naive Bayes 분류자[12]를 사용한다.

Naive Bayes 분류자는 학습 단계와 분류단계를 통하여 문서에 나타나는 모든 단어를 특징으로 분류한다. 실험 문서(D)의 특징이 $\{n_1, n_2, \ldots n_k, n_m\}$라고 하였을 경우 식(1)에 의해서 {class1, class2, classID, classN}중 하나의 클래스로 실험 문서(D)를 분류한다.

$$P(class_{ID}|D) = \frac{P(class_{ID})}{P(D)} \prod_{k=1}^{|D|} P(n_k|class_{ID}) \quad (1)$$

식(1)에서 P(classID)는 classID로 분류될 확률이며, P(nk|classID)는 n_k가 classID에 포함될 확률이다. $\{n_1, n_2, \ldots n_k, n_m\}$의 각 단어는 문맥이나 위치에 관계없이 독립적이라고 가정한다. 독립 가정을 전제로 하는 각 단어에 대한 $P(n_k|class_{ID})$의 확률은 식(2)에 대입함으로써 구한다.

$$P(n_k|class_{ID}) = \frac{n_{kclass_{ID}} + 1}{n_{class_{ID}} + |V|} \quad (2)$$

식(2)에서 $nclass_{ID}$는 $class_{ID}$내의 단어의 총 개수이며, $nkclass_{ID}$는 $class_{ID}$에서 단어 nk의 출현 빈도수, 그리고 $|V|$는 $class_{ID}$의 총 어휘수이다. 분자에 $nkclass_{ID}$에 1을 더하여 확률이 0이 되는 것을 예방하는 Laplace smoothing[13]방법을 사용한다.

이러한 경우, 영화를 장르별로 분류하기 위해 학습 집단을 사용하여 식(3)에 의해 영화를 장르별로 분류할 수 있다.

$$P(class_{ID}|M) = \frac{P(class_{ID})}{P(M)} \prod_{m=1}^{S} \prod_{i=1}^{|d_m|} P(a_{mi}|class_{ID}, s_m) \quad (3)$$

여기서 d_m은 영화를 문서의 벡터로써 표현한다. s_m은 m번째 슬롯을 나타내고, S는 슬롯의 수를 나타낸다. a_{mi}는 m번째 슬롯의 i번째 단어를 나타낸다. $P(n_k|class_{ID}, s_m)$은 $class_{ID}$와 슬롯에서의 각각의 단어들의 확률 값을 나타낸다. 슬롯에 포함될 확률의 값이 가장 높은 것이 추천이 되는 대상이다.

2.2 협력적 여과 시스템

협력적 여과 시스템은 사용자의 선호도에 대한 데이타를 기반으로 사용자가 관심을 가질 것으로 생각되는 아이템을 추천해주는 기법이다.

협력적 여과에서 가장 우선적으로 필요한 것은 특정 사용자와 유사한 선호도를 가지는 이웃을 찾아 내는 것이다. 유사도 가중치 값이 관계없이 유사도 가중치가 구해진 모든 이웃들을 사용해서 선호도를 예측할 수 있지만 이는 성능이나 정확도면에서 그리 좋은 방법은 아니다. 반면 너무 유사도가 높은 이웃들만을 예측해서 고려할 경우 다른 고객들과 유사도가 높지 않은 고객의 아이템에 대해서 예측을 할 수 없는 경우가 발생한다. 그러므로 시스템이 예측을 할 수 있는 적절한 이웃의 수를 결정하는 것이 무엇보다도 중요하다. 예측에 사용될 이웃의 수를 결정하기 위해서 Thresholding과 Best-n-

neighborhood을 사용한다[3,14].

Thresholding은 사용자간의 유사도 가중치가 어느 정도의 값 이상인 이웃들만 사용해서 예측하도록 제안하는 방법이고, Best n-neighborhood는 특정 사용자와 유사한 n명의 이웃을 사용해서 예측하도록 제안하는 방법이다. 위의 두 방법을 조합하여 유사도 가중치가 어느 정도 값 이상인 이웃들 중 n명을 사용할 수 있도록 하는 것도 좋은 방법이다. 여기서 사용자간의 유사도 가중치를 계산하기 위해서 사용되는 대표적인 유사도 기준 값으로 Pearson correlation coefficient가 사용된다.

기본적으로 협력적 여과 시스템의 알고리즘은 4가지 단계를 거쳐 구현 평가되는데 그 단계는 아래 표 1과 같다[14,15].

표 1 협력적 여과 시스템의 단계별 기법

단계1. 활성화된 사용자(active user)와 이웃들과의 유사도 가중치 값을 정의하고 계산한다.
단계2. 활성화된 사용자의 특정 아이템에 대한 선호도를 예측하기 위해서 유사도가 높은 이웃을 어떤 기준으로 몇 명을 선택할지를 결정한다.
단계3. 유사한 선호도를 가지는 이웃들의 아이템에 대한 선호도를 기반으로 활성화된 사용자의 선호도가 입력되지 않은 아이템들의 선호도 값을 예측한다.
단계4. 활성화된 사용자가 선호도를 입력하지 않은 아이템들의 실제 선호도와 예측된 선호도를 가지고 협력적 여과의 결과를 적절한 기준으로 평가한다.

단계 1에서 Pearson correlation coefficient를 사용하여 사용자 a와 사용자 i의 유사도 가중치는 식(4)과 같이 정의된다.

$$w(a,i)=\frac{\sum_{j=1}^{m}(v_{a,j}-\overline{v_a})\times(v_{i,j}-\overline{v_i})}{\sqrt{\sum_{j=1}^{m}(v_{a,j}-\overline{v_a})^2\times\sum_{j=1}^{m}(v_{i,j}-\overline{v_i})^2}} \quad (4)$$

$v_{a,j}$ 는 사용자 a가 아이템 j에 대해서 보여준 선호도이고, 는 사용자 a가 선호도를 입력한 아이템들에 대한 선호도 평균값이다. j는 사용자 a와 i가 공통으로 선호도를 입력한 아이템들이고, m은 아이템의 총 개수이다.

단계 3에서는 식(5)의 Deviation-from-mean 방법을 사용하여 예측 선호도 값을 계산한다.

$$p_{a,k}=\overline{v_a}+\frac{\sum_{i=1}^{n}w(a,i)\times(v_{i,k}-\overline{v_i})}{\sum_{i=1}^{n}w(a,i)} \quad (5)$$

$P_{a,k}$는 사용자 a의 아이템 k에 대해서 선호도를 예측한 값이고, $\overline{v_a}$는 사용자 a의 선호도 평균값이다. $w(a,i)$는 사용자 a와 사용자 i의 유사도 가중치이고, n은 단계 2에서 결정된 이웃들 안의 사용자의 수이다.

2.3 유사도 가중치에 대한 다른 고려 사항

2.3.1 Significance weighting

사용자의 유사도 가중치를 구하는 과정에서 사용자 a와 사용자 i가 공통으로 선호도를 입력한 아이템의 개수가 적은 경우에 사용자 간의 유사도 가중치 값이 매우 높게 나오는 경우가 많이 발생한다. 이러한 경우 특정 사용자의 아이템에 대한 선호도를 예측했을 경우 정확도가 떨어 지는 경우가 많이 발생하기 때문에 사용자간의 유사도 가중치를 구할 때 사용자들이 공통으로 선호도를 입력한 아이템 개수의 제한을 정하고 제한된 개수에 미치지 못하는 사용자간의 유사도 가중치에는 Significance weighting값을 부여할 수 있다. 공통으로 선호도를 입력한 아이템 개수 제한이 m이고 m보다 작은 공통으로 선호도를 입력 아이템 개수 n을 가지는 사용자 a와 사용자 i사이의 유사도 가중치에는 n/m의 Significance weight를 곱해준다. n이 m보다 클 경우에는 Significance weight를 1로 준다[14]. 예를 들어 두 명의 사용자가 공통으로 입력한 아이템의 개수가 45개 미만일 경우 Significance weight $sig_{a,u}$ = n/45을 두 사용자의 유사도 가중치를 계산하는 부분에 적용한다. 공통으로 입력한 아이템의 개수가 45개 이상인 경우 $sig_{a,u}$ =1을 적용한다.

사용자간에 공통으로 선호도를 보인 아이템 개수가 개수의 제한보다 작은 경우 유사도 가중치가 정확하지 않을 수 있지만 이러한 경우에 대상 사용자에서 제외할 경우 예측이 가능한 아이템의 개수를 제한할 수도 있으므로 일반적으로 Significance weight를 적용하는 방법을 사용한다.

2.3.2 Variance weighting

사용자의 유사도 가중치를 구하는 과정에서 고려해 볼 수 있는 Variance weighting이다. 영화에 대한 선호도를 예를 들어 보면, 대부분의 영화사이트 사용자들은 '타이타닉' 이라는 영화에 대해서 높은 선호도를 보였다. 결국 '타이타닉' 이라는 영화는 사용자간의 유사도 차이에 크게 영향을 미치지 못한다. 반면 '러브 스토리'이라는 영화에 대해서 액션 영화를 좋아하는 사용자와 로맨스 영화를 좋아하는 사용자가 상이한 선호도를 보일 수 있다. 이 경우 '러브 스토리'이라는 영화는 '타이타닉' 보다는 사용자간의 선호도 차이에 대해서 더 많은 정보를 가지며 이는 '러브 스토리'라는 영화에 대해 사용자들이 입력한 선호도 분산이 '타이타닉' 이라는 영화에 대해서 사용자들이 입력한 선호도의 분산보다 크다는 것을 의미한다.

위와 같이 사용자를 구분하는 것에 영향력이 큰 아

이템에 대해서 영향력의 강도 값으로 아이템의 분산 값을 적용하는데 사용자 a와 사용자 i의 유사도 가중치인 Pearson correlation coefficient에 Variance weight를 적용하면 식(6)과 같다.

$$w(a,i)=\frac{\sum_j Var_j \times z_{a,j} \times z_{i,j}}{\sum_j Var_j} \quad (6)$$

단, z_{aj}는 를 평균 0, 표준편차 1로 변환한 값이다.

$$Var_j=\frac{Variance_j - Variance_{\min}}{Variance_{\max}}$$

$$Variance_j=\frac{\sum_{k=1}^{n}(v_{k,j}-\overline{v_j})^2}{n-1}$$

n은 아이템 j에 선호도를 보인 사용자의 총 개수이다. 그리고 Varicancemax, Varicancemin은 모든 아이템에 대한 분산 중 최대, 최소 분산을 의미한다.

2.3.3 Case amplification

여러 가지 방법으로 구해진 사용자 a와 사용자 i간의 유사도 가중치의 값은 Significance weighting, Variance weighting을 반영하여 더 정확하고 의미있게 재계산되어 진다. 구해진 유사도 가중치에 대해서 Case amplification 방법은 1에 가까운 유사도 가중치를 더 강조하게 된다. 유사도 가중치의 값에 대해 Case amplification하는 방법은 식(7)과 같다.

$$w'(a,i)=\begin{pmatrix} w^{\rho}_{a,i} \text{if } w_{a,i}\geq 0 \\ -(-w^{\rho}_{a,i}) \text{if } w_{a,i}<0 \end{pmatrix} \quad (7)$$

즉, 적절한 ρ값에 따라 절대값이 1에 가까운 유사도 가중치를 좀더 강조할 수 있도록 유사도 가중치의 값을 조절하는 방법이다[14].

3. 내용 기반 여과와 협력적 여과의 병합을 통한 조화 평균 가중치

3.1 시스템 구성도

그림 1은 내용 기반 여과와 협력적 여과의 병합을 통한 추천 시스템에서의 조화 평균 가중치를 구하여 사용자의 선호도를 예측하는 시스템에 대한 구성도로서 세부 단계의 작업 과정은 다음과 같다.

협력적 여과 시스템의 단점 중 사용자가 평가 하지 않은 아이템들은 사용자에게 추천되지 않는 초기 평가 문제를 해결하기 위해서 내용 기반 여과의 방법을 사용하였다. 희박성 문제를 해결하기 위해서 나이와 성별을 적용한 Representative Attribute-Neighborhood (RA-Neighborhood) 방법[16]을 사용하여 사용자-아이템의 행렬의 차원 수를 줄임으로써 문제를 해결하였다. 아이템의 속성을 고려하지 않는 단점을 보완하기 위해서는

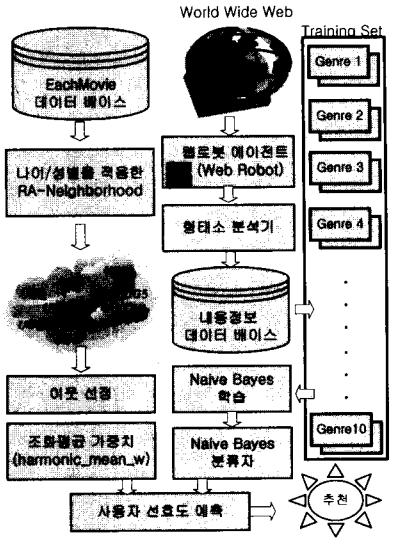

그림 1 시스템 구성도

대표 속성을 중심으로 특정 사용자와 유사한 선호도를 가지는 이웃을 찾아 내는 것이다. 본 논문에서는 사용자간의 유사도 가중치를 구할 때 고려할 요인으로 조화평균 가중치를 제안하여 적용한다.

3.2 내용 정보 데이타베이스 구축

본 논문에서는 문서 표현 형태로 상용으로 사용되는 Bag-of-words[18,19]의 형태를 채택한다. 영화에 대한 정보를 얻기 위해서 웹 로봇 에이전트를 사용하여 영화 관련 URL에서 웹 문서를 수집한다. Bag-of-words의 형태에서는 형태소 분석의 결과로부터 불용어를 제거하고 명사를 추출하여 추출된 명사로 구성된다. 이에 따라 문서내의 단어의 순서와 문서의 구조는 고려되지 않는다. 본 논문에서는 내용정보 데이타 베이스를 구축하기 위해서 각각의 슬롯을 Bag-of-words의 형태로 사용한다.

3.2.1 웹 로봇 에이전트(Web Robot)

그림 1에서 보면 EachMovie 데이타[20]에서 제공하는 영화에 대한 사용자의 평가에 대해서 영화의 정보를 얻기 위해서 영화 관련 URL을 가지고 웹 문서를 수집한다. 영화 관련 웹 문서를 수집할 때 웹 로봇 에이전트를 사용한다. 웹 로봇 에이전트는 웹 상에 존재하는 수천개의 문서를 자동으로 수집한다. 그림 2의 웹 로봇은 수집된 문서들을 가지고 문서의 수와 내용, 제목, URL 등이 저장된 엔트리 파일을 생성하게 된다.

인터넷 서버 응용 프로그램은 HTTP서버가 로드 할 때 연결되는 로드 시간 동적 연결방법(load-time dynamic

linking)과 클라이언트로부터 요구가 있을 때 HTTP서버가 해당 인터넷 서버 응용프로그램을 실행할 때 연결되는 동적 연결 방법(runtime dynamic linking)의 두 가지가 있다.

본 논문에서는 실행 시간 동적 연결 방법을 사용한다. 인터넷 서버 응용프로그램의 동작은 HTTP서버가 실행할 때 ISA의 엔트리 포인트를 호출함으로써 이루어진다. 실행화일 형태의 CGI와는 달리 ISA DLL은 호출한 HTTP서버의 주소 공간에 매핑되고 같은 프로세스에서 동작한다. 이것은 곧 HTTP서버의 자원을 ISA DLL로 공유하여 사용할 수 있다는 의미이다.

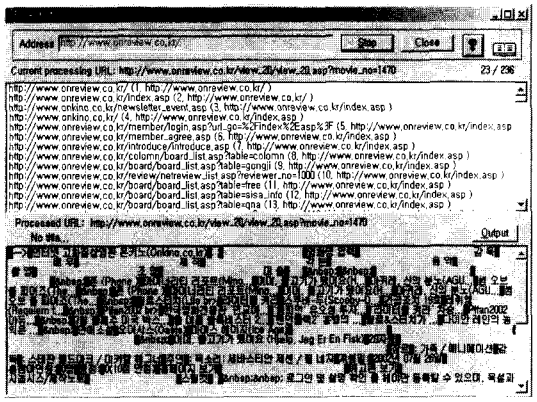

그림 2 웹 문서를 수집하는 웹 로봇 에이전트

그림 2의 화면 상단에 있는 Address란에 URL을 넣으면 그 도메인 내에 있는 홈페이지들을 수집하여 홈페이지의 태그들을 제거하고 홈페이지 내용과 제목을 저장한다. 그러면 해당 URL은 그림 2의 화면 하단에 나타나게 된다. 본 논문에서는 영화 관련 홈페이지를 대상으로 실험을 하였으며 수집된 웹 문서의 개수는 4485개이다. 이를 전처리 과정[21]을 거쳐 1628개의 웹 문서를 가지고 실험을 진행하였다.

3.2.2 형태소 분석기

웹 로봇 에이전트는 웹 상에 존재하는 수천개의 웹 문서들을 수집한 후 영화에 대해서 각각의 슬롯을 구성하기 위해 형태소 분석을 통한 명사 추출 과정을 전처리 과정으로 사용한다.

정보 검색에서 형태소 분석은 검색 시스템의 검색 효율을 높이기 위해서 문서에서 적절한 색인어를 추출해야 하는데 이를 위해서 불필요한 단어는 제거해야 한다. 영어권의 검색 시스템은 불필요한 단어를 여과하는 방법으로 스테밍(Stemming) 방법을 사용한다. 한글에서는 불필요한 단어를 추출하기 위해서 형태소 분석이 필요하다. 형태소 분석에 사용되는 시스템은 사용자 중심의 지능형 정보검색 시스템[23]에서 사용된 방법이다. 이 방법에서는 전처리 단계로 어절분리를 하고 준말을 본디말로 환원한다. 1단계에서는 한어절의 가능한 모든 형태소 결합을 형성한다. 2단계에서는 1차 필터링과 불규칙처리를 하며 3단계에서는 2차 필터링과 후편집을 행하는 과정을 거쳐 형태소 분석이 이루어진다. 마지막으로 형태소 분석의 복잡한 부분인 파싱(parsing)을 통한 의미 분석을 생략하여 표 2의 형태소 분석기의 품사 분류 체계[21,22]에 의해서 결과가 나온다. 본 논문에서는 웹 문서의 형태소 분석의 결과 여러 품사 분류 체계 중에 명사를 기반으로 보통명사(N), 고유명사(Np), 의존명사(Nd)를 사용한다.

표 2 형태소 분석기의 품사 분류 체계

체언	명사	보통명사(N)
		고유명사(Np)
		의존명사(Nd)
	대명사(PN)	
	수사(NU)	
수식언	관형사(D)	
	부사(AD)	
독립언	감탄사(G)	
관계언	조사	격조사(Pc)
		접속조사(Pi)
		보조사(Ps)
용언	동사(V)	
	형용사(A)	
	지정사(IDA)	
부속어	어미	선어말어미(Epf)
		어말어미(En)
	접사	접미사(SF)
	매개모음(EU)	

형태소 분석의 결과를 바탕으로 명사만을 대상으로 8개의 슬롯을 구성한다. 각 슬롯은 영화명, 장르, 감독, 제작, 각본, 음악, 주연, 줄거리로 구성한다. 각 영화를 홍보하는 웹 문서를 각각의 슬롯을 기반으로 내용 정보 데이타 베이스를 생성한다.

3.3 Representative Attribute-Neighborhood

아이템의 속성을 고려하지 않는 협력적 여과 시스템의 단점으로 보완하여 좀더 효율적인 여과를 수행하기 위해 대표 속성을 중심으로 특정 사용자와 유사한 선호도를 가지는 이웃을 찾아 내는 것이다. 여기서 대표 속성이란 선호도에 가장 크게 영향을 미치는 속성을 의미한다. 기존의 협력적 여과 방법은 사용자의 각 정보에 대한 선호도의 정도를 반영하여 예측을 수행하기 위해 전체 정보에 대하여 유사도를 계산하여 예측에 반영하게 된다. 그러나 전체 정

보를 모두 사용하여 유사도를 구하는 것은 대표 장르에 대해서 사용자가 차별적인 선호도를 가지는 경우 이를 제대로 반영하지 못하는 단점이 있다. 그러므로 본 논문에서는 이를 보완하기 위해서 사용자의 대표 장르 추출 알고리즘에 의해 대표 장르에 한정하여 유사한 이웃을 찾아 낼 때 이를 예측에 이용하는 Representative Attribute-Neighborhood(RA-Neighborhood)[16]를 사용한다. 그림 3은 성별과 나이를 적용한 RA-Neighborhood 개념도이다.

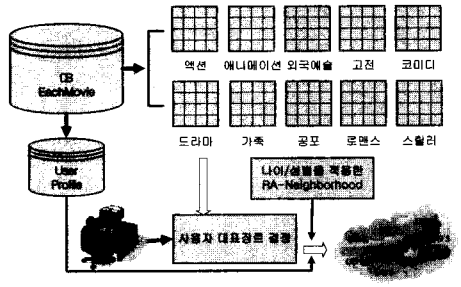

그림 3 성별과 나이를 적용한 RA-Neighborhood 개념도

사용자가 선호도를 보인 아이템을 사용하여 사용자의 대표장르를 구한다. 대표 장르를 추출하기 위해서는 사용자의 장르별 아이템의 선호도 합을 구한 후 선호도 합이 가장 큰 장르를 대표 장르로 정한다[17]. 대표 장르를 추출하는 이유는 실험 데이타로 쓰이는 EachMovie 데이타에서 영화에 대한 장르가 하나 이상인 것이 많이 있기 때문이다. 예를 들어 '타이타닉'의 장르는 드라마, 로맨스, 가족의 3가지 장르를 가지고 있다. 여기서 이 영화는 사용자의 관점에 따라 영화의 대표 장르는 달라지기 때문에 대표 장르를 추출해야 한다[16]. 알고리즘 1은 사용자의 대표장르를 추출하는 알고리즘이다.

```
Num_class ← # of item in GenreID;
MainGenreID ← Null;
MainGenreMaxSum ← 0;
For(j=1; j ≤ Num_class; j++){
  GenreMaxSum ← 0;
  For(each item){
     GenreMaxSum ← GenreMaxSum + Score;
} // 아이템에 대해서 장르별 선호도의 합을 구한다.
If (GenreMaxSum > MainGenreMaxSum){
        MainGenreID ← GenreID of j th;
        MainGenreMaxSum ← GenreMaxSum;
} // 선호도의 합이 가장 큰 장르의 ID을 Return
}
Assign(MainGenreID); // 대표장르 결정
MainGenre-Neighbor[MainGenreID] ← Add UserID;
// Representative Attribute-Neighborhood
```

알고리즘 1 사용자의 대표장르를 추출하는 알고리즘

Representative Attribute-Neighborhood를 사용하여 유사한 이웃을 선택을 할 때 사용자의 성별과 나이를 군집을 사용한다. 본 논문에서는 같은 성별 또는 같은 나이를 가진 사람들은 각각의 대표 속성 값에 대해서 유사한 선호도를 가진다고 가정한다. 표 3은 성별과 나이를 군집하기 위한 히스토 그램이다.

표 3 나이와 성별 군집에 따른 히스토그램

Group		User(n)
Gender	Male	24724
	Female	7107
Age	1-14	916
	15-19	3137
	20-24	4611
	25-29	5422
	30-39	7649
	40-49	6203
	50-59	2623
	60-69	594

표 3에서 본 논문에서 정의한 성별과 나이 군집을 제안하는 알고리즘에 실시간으로 적용을 하면 예측의 정확도는 향상된다. 그러므로 특정 사용자와 유사한 선호도를 가지는 이웃을 찾아 내기 위해서 Thresholding과 Best-n-neighborhood을 사용하는 기존의 방법[14] 대신에 성별과 나이를 적용한 Representative Attribute-Neighborhood 방법을 사용한다. 알고리즘 2는 Representative Attribute-Neighborhood 방법에 성별과 나이를 적용하여 군집하는 알고리즘이다.

```
Num_class ← # of item in GenreID;
Num_gender ← # of item in Gender;
Num_age ← # of item in Age;
For(i=1; j ≤ Num_class; i++){
  For(j=1; j ≤ Num_gender; j++){
    For(k=1; k ≤ Num_age; k++){
      UserGroup(i, j, k) ← 조건 i, j, k를 만족하는 User Group.
    }
  }
}// Representative Attribute-Neighborhood with gender and age group
Assign(User Group);
```

알고리즘 2 성별과 나이를 적용한 RA-Neighborhood

3.4 조화 평균 가중치

본 논문에서 사용자 유사도 가중치를 계산할 때 고려할 요인으로 조화 평균 가중치를 제안한다. 유사도 가중치를 구할 때 사용자들이 평가한 아이템의 개수가

예측 정확도에 중요한 요인이 된다. 예를 들어 사용자가 많은 아이템에 평가를 하면 내용 기반 여과는 좋은 성능을 가진다. 그러나 적은 수의 아이템을 평가를 하면 예측의 정확도가 낮아 진다. 또한 사용자 간의 유사도 가중치 값이 매우 높게 나오는 경우가 발생을 한다. 이러한 경우 특정 사용자의 아이템에 대한 선호도 예측을 했을 때 정확도가 떨어지는 경우가 발생한다. 이를 보완하여 사용자 유사도 가중치를 구할 때 평균 개념을 이용한다. 조화 평균 가중치를 수식으로 유도하기에 앞서 산술 평균 가중치와 기하 평균 가중치로 수식을 유도 한다.

산술 평균 가중치는 식(8)과 같이 나타낼 수 있다.

$$arithmetic\_mean\_w_i = \sum_{i=1}^{m} r_i / m \quad (8)$$

식(8)의 산술 평균 가중치는 모든 Ri(i=1,2,..,m)에 동일한 가중치($1/m$)를 준 것이다. 이 방법은 공통으로 입력한 아이템의 개수에 따라 예측의 정확도가 달라지는 Significance weight에 적용할 수 없다.

기하 평균 가중치는 식(9)와 같이 나타낼 수 있다.

$$geometric\_mean\_w_i = \prod_{i=1}^{m} R_i^{1/m} \quad (9)$$

식(9)의 기하 평균 가중치는 계산 방식이 복잡하여 실시간으로 사용자 유사도 가중치를 구하기에는 부적합하다. 산술 평균 가중치와 기하 평균 가중치의 단점 때문에 본 논문에서는 산술 평균 가중치에 기반을 둔 조화 평균 가중치를 사용한다. 사용자의 유사도 가중치를 구하는 부분에 조화 평균 가중치를 사용한다. 조화 평균 가중치는 식(10)과 같이 정의한다.

$$harmonic\_mean\_w_{i,j} = \frac{1}{\frac{\left(\frac{1}{R_1} + \frac{1}{R_j}\right)}{2}} = \frac{2R_iR_j}{R_i + R_j} \quad (10)$$

$$R_i = \begin{cases} \frac{n_i}{45} : if\ n_i < 45 \\ 1 : ohter \end{cases}$$

식(10)에서 n_i는 사용자 i가 평가한 아이템의 수이다. 조화 평균 가중치는 R_i과 R_j의 값에 대해서 낮은 값으로 가중치가 부여되는 경향이 있다. 그러므로 적어도 45개의 아이템을 평가한 사용자들 사이에서 유사도 가중치를 구할 때 사용자가 실제적으로 평가한 아이템의 수에 관계없이 높은 가중치 값을 가진다. 반면 사용자가 평가한 아이템의 개수가 45개 보다 적을 경우 낮은 가중치의 값이 부여된다. 여기서 사용자 i가 평가한 아이템의 수를 45개로 임계치를 설정한 것은 그림 4의 내용 기반 여과의 학습 곡선에 기반을 둔 것이다.

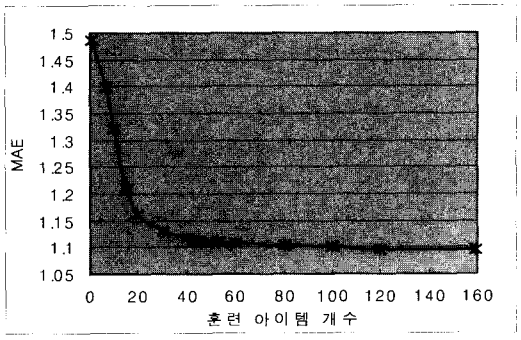

그림 4 내용 기반 여과에서 학습 곡선

그림 4를 보면 훈련 아이템의 개수가 늘어날수록 예측의 성능이 좋아진다. 그러나 훈련 아이템의 개수가 45개인 부분에서 내용 기반 여과의 학습 곡선이 완만해지는 것을 볼 수 있다. 그러므로 본 논문에서는 실험을 통해 내용 기반 여과의 예측의 성능을 고려하여 임계치를 45으로 설정한다. 이는 훈련 집합의 크기에 관계없이 예측의 정확도가 유지되기 때문이다.

그림 5는 산술 평균 가중치, 기하 평균 가중치, 조화 평균 가중치를 유사도 가중치를 구할 때 고려할 요인으로 적용하였을 경우의 정확도를 보인다.

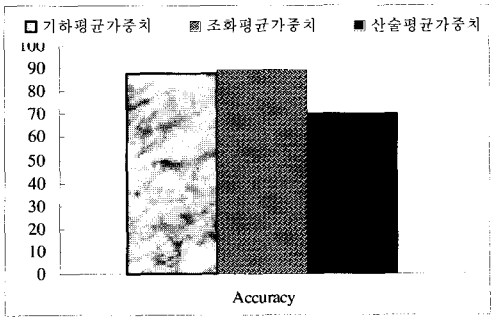

그림 5 기하 평균 가중치, 조화 평균 가중치, 산술 평균 가중치를 적용했을 때의 정확도

그림 5에서와 같이 조화 평균 가중치를 적용을 할 경우의 정확도는 89.3으로 기하평균 가중치보다는 4.07%, 산술 평균 가중치보다는 13.87% 높다. 그러므로 사용자의 유사도 가중치를 계산할 때 조화 평균 가중치를 적용하는 것이 가장 바람직하다.

조화 평균 가중치를 구하기 위해 Significance weight을 적용하면 다음 식(11)과 같이 혼합된 상관 관계 가중치(hybridwa,u)를 구할 수 있다.

$$Hybridw_{a,u} = harmonic\_mean\_w_{a,u} + sig_{a,u} \quad (11)$$

3.5 내용 기반 여과와 협력적 여과의 병합을 통한 사용자의 선호도 예측

내용 기반 여과와 협력적 여과의 병합을 통한 추천 시스템에서의 사용자 유사도 가중치를 구할 때 조화 평균 가중치(CBCF_harmonic_mean)를 적용한다. 사용자 a가 아이템 i에 대해서 내용 기반 여과와 협력적 여과(CBCF)에서의 예측은 식(12)과 같이 정의한다.

$$p_{a,k} = \overline{v_a} + \frac{(c_{a,k} - \overline{v_a}) + \sum_{\substack{i=1 \\ u \neq a}}^{n} Hybridw_{a,u} w(a,i)(v_{u,i} - \overline{v_u})}{\sum_{\substack{i=1 \\ u \neq a}}^{n} Hybridw_{a,u} w(a,i)} \quad (12)$$

식(12)에서 $P_{a,k}$는 사용자 a의 아이템 k에 대해서 선호도를 예측한 값이고, $C_{a,k}$은 사용자 a가 아이템 k에 대해서 내용 기반 여과 시스템에서 예측한 값이다. 이는 내용 정보 테이타 베이스에서 슬롯에서의 각각의 단어들 중 확률 값이 가장 높은 것이 추천의 대상이 된다. Hybridwa.k는 조화 평균 가중치에 $n/45$의 Significance Weight을 사용한 혼합된 상관관계 가중치이다. $\overline{v_a}$는 사용자 a의 선호도 평균값이다. $w(a,i)$는 사용자 a와 사용자 i의 유사도 가중치이고, n은 성별과 나이를 적용한 Representative Attribute-Neighborhood에 의해 군집된 사용자의 수이다.

4. 실험 및 결과

4.1 실험 환경 및 실험 데이타

본 논문에서 제안한 내용 기반 여과와 협력적 여과의 병합을 통한 추천 시스템에서의 사용자 유사도 가중치는 Microsoft Visual C++ 6.0으로 구현되었으며, 실제 실험 환경은 PentiumⅢ 450Mhz, 256MB RAM 환경에서 수행되었다. 실험 데이타로는 컴팩 연구소에서 18개월 동안 협력적 여과 알고리즘을 연구하기 위해서 영화에 대한 사용자의 선호도를 조사한 EachMovie 데이타[20]를 사용한다. 이 데이타에는 영화에 관한 사용자의 선호도가 0, 0.2, 0.4, 0.6, 0.8, 1.0의 6단계의 수치로 표현되어 있다.

내용 기반의 여과의 추천을 위한 데이타베이스는 1628개의 서로 다른 웹 문서로 구성한다. 1628개의 웹 문서는 웹 문서 수집기에 의해서 영화 관련 분야의 URL로부터 수집된 문서이다. 1628개의 훈련 문서는 내용 정보 데이타 베이스를 생성하기 위한 10개의 장르로 분류한다. 여기서 10개의 클래스는 {액션, 애니메이션, 외국 예술, 고전, 코미디, 드라마, 가족, 공포, 로맨스, 스릴러}의 레이블이다. 10개의 장르 클래스로 분류한 기준은 EachMovie 데이타에서 제공해 주는 장르에 따른 것이다. 협력적 여과의 추천을 위한 데이타베이스는 최소 100회 이상 선호도를 입력한 사용자 4798명을 추출하여 이 가운데 1000명을 기존 사용자 군으로 두고 나머지 사용자들 중에 무작위로 테스트 사용자 100명을 선택하여 총 1628개의 영화 중 테스트 사용자가 선호도를 표시한 임의의 10개의 영화에 대하여 선호도를 예측하고 실제 선호도와 비교하였다. 대표 장르를 추출할 때 사용자의 대표 속성은 10개의 장르 중에 하나로 가정하고 실험을 진행 하였다.

4.2 웹 문서로부터 내용 정보를 추출한 예

본 절에서는 영화 홈페이지에서 영화에 대한 내용 정보를 웹 로봇 에이전트(Web Robot)에 의해 수집하여 형태소 분석기를 통한 내용 정보를 추출하는 절차를 보인다. 그림 6은 내용 정보를 추출하기 위한 웹 문서이다. 그림 6의 웹 문서는 영화에 해당하는 내용 정보에 대한 홍보를 하는 홈페이지이다.

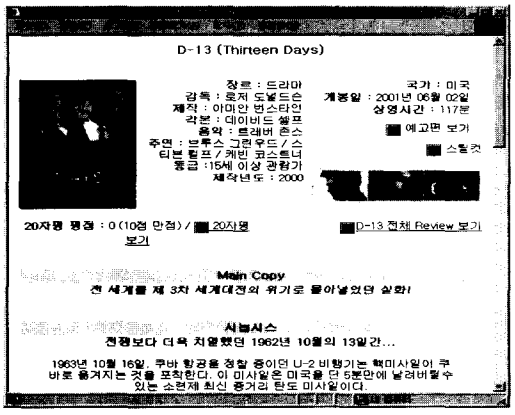

그림 6 http://www.onreview.co.kr/movie_info/movie_synop.asp?movie_no=1119의 URL에 나타난 웹문서

그림 7는 그림 6의 웹 문서를 형태소 분석한 결과의 일부이다. 형태소 분석에 사용되는 시스템은 사용자 중심의 지능형 정보 검색 시스템[22,23]에서 사용된 방법이다. 웹 문서 안에 있는 이미지(JPG, BMP, GIF), Hyperlink (URL), 수치 정보(날짜 정보, 시간 정보)는 형태소 분석을 할 때 제외된다.

형태소 분석의 결과를 보면 웹 문서에 대한 품사가 분류{보통명사(N), 고유명사(Np), 의존명사(Nd), 대명사(PN), 수사(NU), 관형사(D), 부사(AD), 감탄사(G), 격조사(Pc), 접속조사(Pi), 보조사(Ps), 동사(V), 형용사

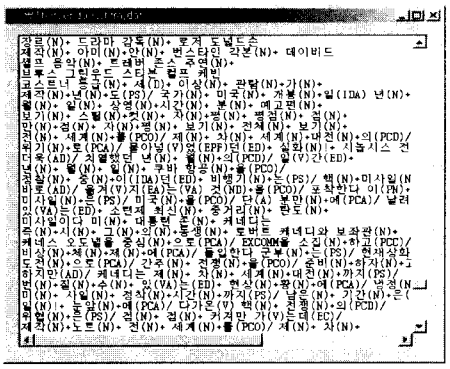

그림 7 웹 문서를 형태소 분석한 결과

(A), 지정사(IDA), 선어말어미(Epf), 어말어미(En), 접미사(SF))가 되어 있는 것을 볼 수 있다.

본 논문에서는 추출된 보통명사(N), 고유명사(Np), 의존명사(Nd)를 기반으로 8개의 슬롯을 구성한다. 각 슬롯은 영화명, 장르, 감독, 제작, 각본, 음악, 주연, 줄거리로 구성한다. 각 영화를 홍보하는 웹 문서를 각각의 슬롯을 기반으로 표 4의 내용 정보 데이타베이스를 생성한다.

표 4 13-Days에 대한 내용 정보 데이타베이스

슬롯	추출된 명사
영화명	Thirteen Days
장르	드라마
감독	로저 도널드슨
제작	아미안 번스타인
각본	데이비드 셀프
음악	트레버 존스
주연	브루스 그린우드, 스티븐 컬프, 케빈 코스트너
줄거리	세계, 대전, 위기, 실화, 쿠바 항공, 정찰, 비행기

표 4의 내용 정보 데이타 베이스는 2.1절에서 언급한 내용 기반 여과를 할 때 쓰인다.

4.3 Representative Attribute-Neighborhood에 의한 사용자 군집

알고리즘 1에 의해서 대표 장르가 결정된 사용자들은 유사한 이웃을 찾아 내어 예측을 하는 Representative Attribute-Neighborhood 방법에 사용된다. 표 5에서 대표 장르가 결정된 사용자들의 많은 수가 액션 장르와 드라마 장르로 결정 되었다. 이는 대부분의 사용자들이 이 두 장르를 선호하기 때문이다.

대표 장르가 추출된 사용자들을 기반으로 RA-Neighborhood을 만든다. 각 장르별 사용자들을 알고리즘 2의 성별과 나이로 사용자들을 군집 시킨다. 표 6은 표 5의 대표 장르가 추출된 사용자들 중 액션 장르에

표 5 대표 장르가 추출된 사용자들

장르	User ID	User(n)
액션	User9, User10, User11, User25, User26	13590
애니메이션	User7, User8, User12, User24, User27	125
외국예술	User19 User6, User21, User22, User23	385
고전	User1, User4, User20, User35, User36	249
코미디	User2, User10, User33, User34, User48	4107
드라마	User3, User 16, User23, User39, User49	11559
가족	User13, User14, User17, User40, User46	158
공포	User5, User28, User31, User41, User42	74
로맨스	User19, User29, User32, User43, User44	166
스릴러	User15, User30, User37, User38, User45	448

속한 사용자들을 대상으로 성별과 나이를 적용한 RA-Neighborhood을 이용하여 사용자 군집을 만든다.

표 6 액션장르에 성별과 나이를 적용한 RA-Neighborhood

Num_class = Action		User(n)	
		Male	Female
Age	1-14	295	85
	15-19	1227	278
	20-24	1915	349
	25-29	2052	368
	30-39	2676	547
	40-49	2032	434
	50-59	774	131
	60-69	137	26

4.4 RA-Neighborhood의 이웃들의 선택

성별과 나이를 적용한 RA-Neighborhood에 의한 사용자 군집은 적절한 이웃의 수를 결정해야 한다.

실험을 통해서 적절한 이웃의 수를 결정하기 위해서 이웃들의 수를 증가시킴에 따른 정확도를 비교 평가하였다. 그림 8은 RA-Neighborhood의 이웃들의 수를 선택하는 그림이다.

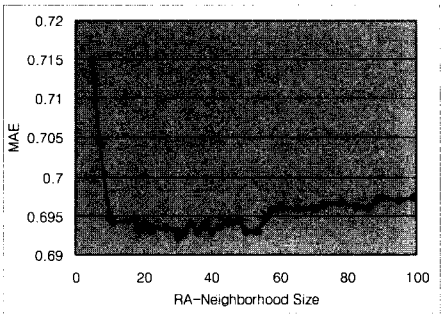

그림 8 RA-Neighborhood의 이웃들의 수 선택

그림 8에서 보면 RA-Neighborhood의 이웃들의 수가 증가함에 따라 정확도가 일관성 있게 좋아지지 않는다. 대략 이웃의 수가 50 정도에 해당되는 곳에서부터 정확도가 감소되는 것을 볼 수 있다. 위의 RA-Neighborhood의 이웃들의 수를 선택하는 실험은 Pearson correlation coefficient로 사용자 유사도 가중치를 구할 때 n/45의 Significance weight를 적용한 것이다.

4.5 분석 및 평가

4.5.1 성능 평가 기준

추천의 성능을 평가 하기 위해 본 논문에서는 Breese [14]에 의해 제안된 순위 스코어 측정(Rank scoring metric)와 MAE(Mean Absolute Error)를 사용한다.

순위 스코어 측정은 순위가 있는 목록에 있는 아이템을 사용자가 평가하는 가의 측정이다. 순위 스코어 측정은 아이템을 선택할 확률이 목록의 하단으로 갈수록 지수적으로 감소한다는 전제에서 측정된다. 각 아이템은 사용자 선호도의 가중치의 값에 따라 내림차순으로 j에 의해 정렬되어 있다고 가정한다. 식(13)은 순위가 부여된 아이템의 목록에 대한 사용자 U_a의 순위 스코어 측정에 대한 기대 이용도(Expected utility)를 계산하기 위한 식이다.

$$R_a = \sum_j \frac{\max(V_{a,j} - d, 0}{2^{(j-1)/(\alpha-1)}} \quad (13)$$

식(13)에서 d는 아이템에 대한 중간 평가 값이며 α는 반감기(halflife)이다. 반감기는 사용자가 평가하거나 방문할 50-50의 기회가 있는 목록에 있는 아이템의 수이다. 본 논문의 평가에서는 반감기를 5로 사용한다. 식(14)는 순위 스코어 척도를 사용하여 새로운 사용자에 대한 예측의 정확도를 나타내는 식이다.

$$R = \frac{\sum_{i=1}^{N} R_i}{\sum_{i=1}^{N} \max(R_i)} \times 100 \quad (14)$$

식(14)에서 $\max(R_i)$는 사용자가 평가한 아이템이 순위가 있는 목록상에서 상위에 나타났을 경우에 측정된 순위 스코어 측정에 대한 기대 이용도의 최대값이다.

MAE는 예측의 정확도를 측정하기 위해서 실제로 사용자가 평가한 값과 예측된 값의 차이에 대한 절대값의 평균을 나타낸다. MAE는 절대적으로 알고리즘이 얼마나 정확하게 예측을 했는지를 알 수 있으며 식(15)에 의해 정의된다.

$$S_a = \frac{\sum_{j \in P_a} |P_{a,j} - v_{a,j}|}{m_a} \quad (15)$$

식(15)에서 $p_{a,j}$는 예측된 선호도이며 $v_{a,j}$는 실제로 사용자가 평가한 선호도이다. 또한 m_a는 새로운 사용자에 의해 평가된 상품의 수를 의미한다.

4.5.2 제안한 방법의 성능 평가

본 논문에서는 제안한 방법(CBCF_harmonic_mean)과 내용 기반 여과 시스템과 협력적 여과 시스템을 병합한 기존의 혼합형 추천 시스템 중 희박성 문제를 해결한 Soboroff 방법[11], 희박성과 초기 평가 문제를 해결한 Pazzani 방법[6], 초기 평가 문제를 해결한 Fab 방법[7]과 사용자가 아이템에 대해 평가한 횟수를 변화시켜가면서 순위 스코어 측정과 MAE[14]를 사용하여 비교하였다.

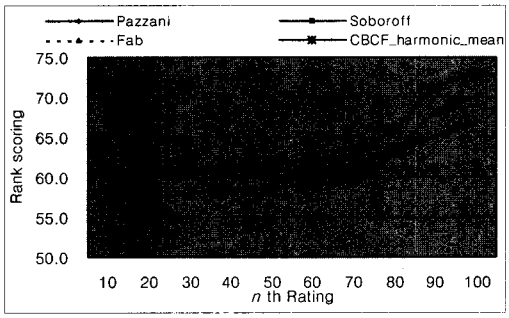

그림 9 n번째 평가에서의 순위 스코어

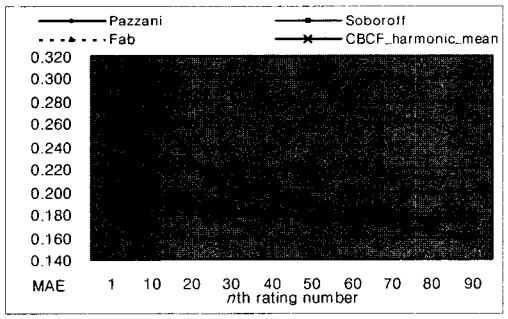

그림 10 n번째 평가에서의 MAE

그림 9와 그림 10은 사용자가 평가한 횟수를 증가시킴에 따른 순위 스코어와 MAE 척도를 나타낸다. 초기 평가 문제를 갖는 Soboroff는 평가의 수가 작을 경우 낮은 성능을 보이며 나머지 방법들은 Soboroff의 방법보다 우세한 성능을 보인다. 초기 평가 문제와 희박성을 해결한 Pazzani 방법과 CBCF_harmonic_mean의 성능은 전반적으로 높은 예측 정확도를 보이며 그 중 CBCF_harmonic_mean 은 가장 높은 정확도를 보인다.

5. 결론 및 향후 연구과제

본 논문에서는 내용 기반 여과와 협력적 여과를 병합한 혼합형 추천 시스템에서의 예측의 정확도를 향상 시키기 위해서 조화 평균 가중치(CBCF_harmonic_mean)를 사용자 유사도 가중치를 구할 때 고려할 요인으로 적용하였다. 내용 기반 여과의 예측의 성능을 고려하여 실험을 통해 임계치 값을 45으로 정하였다. 이는 훈련집합의 크기에 관계없이 예측의 정확도가 유지되는 것을 볼 수 있다. 나이와 성별을 적용한 Representative Attribute-Neighborhood을 사용하여 사용자-아이템의 행렬의 차원수가 감소되므로 희박성 문제를 해결하였다. 아이템의 속성을 고려하지 않는 단점을 보완하기 위해서 대표 속성을 적용하였다. 제안된 방법의 성능을 평가하기 위해 기존의 협력적 여과 시스템과 내용 기반 여과 시스템을 병합한 방법과 비교한 결과 기존의 방법보다 높은 성능을 보였다. 향후 제안된 조화 평균 가중치를 사용자 기반과 아이템 기반의 협력적 여과 시스템에 적용하여 추천 시스템의 정확도를 향상 시킬 것이라 기대한다.

참 고 문 헌

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

[2] M. O'Connor and J. Herlocker, "Clustering Item for Collaborative Filtering," In Proceedings of the ACM SIGIR Workshop on Recommender Systems, Berkeley, CA, 1999.

[3] P. Resnick, et. al., "GroupLens: An Open Architecture for Collaborative Filtering of Netnews," Proc. of ACM CSCW'94 Conference on Computer Supported Cooperative Work, pp. 175-186, 1994.

[4] 정경용, 김진현, 이정현, "연관 사용자 군집과 베이지안 분류를 이용한 사용자 선호도 예측 방법," 제28회 한국정보과학회 추계학술발표 논문집(II), pp. 109-111, 2001.

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

[6] M. J. Pazzani, "A framework for collaborative, content-based and demographic filtering," Artificial Intelligence Review, pp. 393-408, 1999.

[7] M. Balabanovic and Y. Shoham, "Fab: Content-based, collaborative recommendation," Communication of the Association of Computing Machinery, Vol. 40, No. 3, pp. 66-72, 1997.

[8] C. Basu and 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.

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

[10] N. Good, J. B. Schafer and J. A. Konstan, A. Borchers, B. Sarwar, J. Herlocker, and J. Riedl, "Combining collaborative filtering with personal agents for better recommendations," In Proceedings of National Conference on Artificial Intelligence (AAAI-99), pp. 439-446, 1999.

[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] T. Michael, *Maching Learning*, McGraq-Hill, pp. 154-200, 1997.

[13] R. Kohavi, B. Becker, and D. Sommerfield, "Improving simple Bayes" In Proceedings of the Europeam Conference on Machine Learning, 1997.

[14] J. S. Breese and D. Heckerman and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," Proc. of the 14th Conference on Uncertainty in Artificial Intelligence, 1998.

[15] J. Herlocker, J. Konstan, A. Borchers and J. Riedl, "An Algorithm Framework for Performing Collaborative Filtering," In Proceedings of ACM SIGIR'99, 1999.

[16] K. Y. Jung, J. K. Ryu, and J. H. Lee, "A New Collaborative Filtering Method using Representative Attributes-Neighborhood and Bayesian Estimated Value," Proceedings of International Conference on Artificial Intelligence: Las Vegas, USA, June 24-27, 2002.

[17] K. Y. Jung, Y. J. Park, and J. H. Lee, "Integrating User Behavior Model and Collaborative Filtering Methods in Recommender Systems," International Conference on Computer and Information Science, Seoul, Korea, August 8-9, 2002.

[18] K. Y. Jung, J. H. Lee, "Prediction of User Preference in Recommendation System using Association User Clustering and Bayesian Estimated Value," Lecture Notes in Artificial Intelligence 2557, 15th Australian Joint Conference on Artificial Intelligence, December 2-6, 2002.

[19] M. Pazzani, D. Billsus, Learning and Revising User Profiles: The Identification of Interesting Web Sites, Machine Learning 27, Kluwer Academic Publishers, pp. 313-331, 1997.

[20] P. McJones, EachMovie collaborative filtering dataset,URL:http://www.research.digital.com/SRC/eachmovie, 1997.
[21] R. Cooley, et al., "Data Preparation for Mining World Wide Web Browsing Patterns," Knowledge and Information Systems, Vol. 1, NO. 1, 1999.
[22] 정영미, *정보검색론*, 구미무역 출판부, 1993.
[23] 인하대학교, 사용자 중심의 지능형 정보 검색 시스템, 최종 연구 개발 보고서, 정보통신부, 1997.

정 경 용
2000년 인하대학교 전자계산공학과(공학사) 2002년 인하대학교 전자계산공학과(공학석사) 2002년~현재 인하대학교 전자계산공학과 박사과정 2001년~현재 에이플러스전자(주) 선임연구원 2003년~현재 가천길대학 뉴미디어과 겸임교수. 관심분야는: 웹 마이닝, 기계학습, 정보검색, CRM, 협력적 필터링, 자연어처리, 전자상거래

류 중 경
1988년 한국방송대학교 전자계산학과(이학사) 1991년 인하대학교 산업대학원(공학석사) 2002년 인하대학교 전자계산공학과(박사수료) 1983년~1992년 대림산업(주) 전산실 1992년~현재 대림대학 컴퓨터정보과 부교수. 관심분야는 Human Computer Interaction, S/W 아키텍쳐, S/W 품질, S/W Re-engineering, 컴퍼넌트 기반 개발, 웹 서비스

강 운 구
1993년 인하대학교 산업기술대학원 정보공학과 졸업(공학석사) 1999년 인하대학교 전자계산공학과(공학박사) 2000년~현재 이노웹 기술고문(주) 1994년~현재 가천길대학 뉴미디어과 부교수. 관심분야는 분산객체 컴퓨팅, 원격교육, 컴퍼넌트 기반 소프트웨어공학, 멀티미디어, P2P 컴퓨팅, 이동 에이전트

이 정 현
1977년 인하대학교 전자공학과 졸업 1980년 인하대학교 대학원 전자공학과(공학석사) 1988년 인하대학교 대학원 전자공학과(공학박사) 1979년~1981년 한국전자기술연구소 시스템 연구원 1984년~1989년 경기대학교 전자계산학과 교수 1989년~현재 인하대학교 컴퓨터공학부 교수. 관심분야는 자연어처리, HCI, 정보검색, 음성인식, 음성합성, 컴퓨터 구조