협력적 여과와 내용 기반 여과의 병합을 통한 추천 시스템에서의 사용자 선호도 발견

(Discovery of User Preference in Recommendation System through Combining Collaborative Filtering and Content based Filtering)

고 수 정 † 김 진 수 †† 김 태 용 ††† 최 준 혁 †††† 이 정 현 †††††

(Su Jeong Ko) (Jin Su Kim) (Tae Yong Kim) (Jun Hyeog Choi) (Jung Hyun Lee)

요 약 최근의 추천 시스템은 협력적 여과 시스템의 희박성과 초기 평가 문제를 해결하기 위하여 내용 기반 여과 시스템과 협력적 여과 시스템을 병합하는 방법을 사용한다. 협력적 여과 시스템은 부가적인 상품을 예측하기 위해 사용자의 선호도에 대한 데이타베이스를 사용한다. 내용 기반 여과 시스템은 상품의 속성과 사용자의 흥미를 대조함에 의해 아이템을 추천한다. 본 논문에서는 두 가지의 기술을 기계 학습 알고리즘에 응용하고 병합함으로써 사용자의 선호도를 발견하는 방법을 기술한다. 제안된 협력적 여과 방법에서는 유전자 알고리즘을 이용하여 Naive Bayes 분류자에 의해 분류된 아이템을 기반으로 사용자 군집을 생성하며 내용 여과 기법에서는 연관 피드백에 의해 사용자의 흥미를 추출함으로써 사용자의 프로파일을 생성한다. 제안된 방법은 웹문서에 대해 사용자가 평가한 데이타베이스에서 평가되며 기존의 방법보다 높은 성능을 나타냄을 보인다.

Abstract Recent recommender system uses a method of combining collaborative filtering system and content based filtering system in order to solve sparsity and first rater problem in collaborative filtering system. Collaborative filtering systems use a database about user preferences to predict additional topics. Content based filtering systems provide recommendations by matching user interests with topic attributes. In this paper, we describe a method for discovery of user preference through combining two techniques for recommendation that allows the application of machine learning algorithm. The proposed collaborative filtering method clusters user using genetic algorithm based on items categorized by Naive Bayes classifier and the content based filtering method builds user profile through extracting user interest using relevance feedback. We evaluate our method on a large database of user ratings for web document and it significantly outperforms previously proposed methods.

1. 서 론

추천 시스템은 물품 혹은 서비스를 사용자에게 제공하기 위한 시스템[1, 2, 3]으로, 최근에는 협력적 여과 기법을 이용한 추천 방법을 많이 사용한다. 일반적으로 협력적 여과 시스템에서 입력 데이타의 구성은 사용자와 상품의 2차원 행렬로 구성되는데 행(Row)은 사용자, 열(Column)은 상품 목록, 행렬의 값은 사용자가 상품에 대해 제공한 선호도를 나타낸다. 협력적 여과 시스템은 사용자가 상품에 대해 제공한 선호도를 바탕으로 선정된 상품에 대한 예측(Prediction)을 수행하며, 모델 기반의 여과와 메모리 기반의 여과 시스템으로 구분된다[4].

이러한 협력적 추천 시스템은 두 가지의 커다란 단점을 갖는다[5, 6]. 첫째, 대부분의 사용자들은 모든 상품에 대해 평가를 하지 않기 때문에 사용자-상품의 행렬

† 정 회 원 : 인하대학교 전자계산공학과
sujung@nlsun.inha.ac.kr
†† 학생회원 : 인하대학교 전자계산공학과
kjspace@nlsun.inha.ac.kr
††† 정 회 원 : 문경대학 인터넷정보계열 교수
tykim@munkyung.ac.kr
†††† 종신회원 : 김포대학 소프트웨어개발전공 교수
jhchoi@kimpo.ac.kr
††††† 종신회원 : 인하대학교 전자계산공학과 교수
jhlee@inha.ac.kr
논문접수 : 2001년 8월 24일
심사완료 : 2001년 10월 2일

은 매우 희박한 특성을 보인다는 것이다. 이로 인하여 메모리 기반[4]의 협력적 여과 추천에 사용되는 몇십만 명의 사용자와 수천 개의 상품을 행렬로 구성하는데 있어서 저장문제가 발생하며 온라인 추천도 효율적이지 못하게 되는 문제점을 발생시킨다. 둘째, 사용자가 전에 평가를 하지 않은 상품은 사용자에게 추천되지 않는다는 초기 평가 문제이다. [6, 7, 8, 9, 10, 11, 20]은 이러한 협력적 여과 추천 방법의 단점을 해결하기 위하여 협력적 여과와 내용 기반 여과를 병합한 방법을 표 1과 같이 제안하였다. 표 1에서 LSI[11], SVD[9] 분류를 사용한 방법은 데이타 입력 차원수를 줄임으로써 협력적 여과의 희박성 문제를 해결 수 있었으나 초기 평가 문제는 해결하기가 어려웠으며, [5, 7, 8, 10]의 방법들은 초기 평가 문제는 해결하였으나 희박성 문제를 해결하지 못하였다. 반면, [6]의 방법에서는 희박성 문제와 초기 평가 문제를 동시에 해결하려는 시도를 하였다.

본 논문에서는 협력적 여과 기법에서의 사용자-상품

표 1 내용 기반 여과와 협력적 여과를 병합한 기존의 추천 방법

제 안	방 법
LSI(Latent Semantic Indexing)를 이용한 여과 방법(Soboroff et al.)[11]	-입력 데이타의 차원을 축소하여 축소된 차원에서 협력적 여과에 응용 -상품 개수 혹은 고객 수가 많을 경우 이점을 나타냄
SVD(Singular Value Decomposition)를 사용(Billsus and Pazzani)[9]	-다차원적인 상품을 분류함으로써 축소된 상품 행렬 차원에서 협력적 여과에 응용 -사용자에게 흥미로운 하나의 상품에 관련 상품 까지 추천
내용 프로파일의 행렬에 협력적 여과를 적용(Pazzani)[6]	-사용자 프로파일을 훈련 집합으로부터 추출된 가중치가 있는 단어의 집합으로 표현 -내용-프로파일 행렬은 여러 사용자들의 프로파일의 모임
분류를 사용(Basu et al.)[8]	-연역 논리 프로그램인 Ripper을 사용하여 사용자와 상품과의 관계를 학습 =>사용자가 상품을 좋아할 것인가 아닌가를 예측
사용자의 선호도를 학습(Lee)[5]	-비슷한 사용자의 평가뿐 아니라 상품 내용에 대한 사용자의 선호도를 학습
사용자의 연관 피드백과 주제 여과를 통한 내용 기반 여과(Fab)[7]	-내용 기반 여과에서 문서는 주제 여과에 의해 여과되어 문서에 대한 순위 목록을 만들며 생성된 목록에 대해 사용자는 연관 피드백을 제공 -내용 기반 여과를 협력적 여과에 응용
개인화 정보 여과 에이전트(Good et al)[10]	-새로운 사용자에 대한 예측은 새로운 사용자의 개인화 에이전트를 협력적 여과에 응용함으로써 이루어짐

행렬의 차원수를 감소시키기 위하여 베이지안 분류를 사용하여 상품을 클래스별로 분류하며, 사용자를 대상으로는 유전자 알고리즘을 사용하여 분류된 상품을 기반으로 사용자를 군집시킨다. 또한, 연관 피드백을 통하여 새로운 사용자의 프로파일을 생성함으로써 초기 평가 문제를 해결한다. 사용자의 선호도는 차원수가 감소된 협력적 여과를 통한 그룹과 연관 피드백을 통해 생성된 사용자 프로파일의 병합을 통하여 발견된다. 상품을 분류하고 사용자에 대한 프로파일을 생성하기 위한 특징 추출의 방법은 Apriori 알고리즘을 사용한 연관 단어를 마이닝 방법을 사용한다.

제안된 방법은 웹문서에 대해 사용자가 평가한 데이타베이스에서 평가되며 기존의 방법보다 높은 성능을 나타냄을 보인다.

2. 내용 기반 여과

본 논문에서는 초기 평가 문제점을 해결하기 위하여 사용자의 연관 피드백을 통한 내용 기반 여과 방법을 제안한다. 제안된 방법은 새로운 사용자가 웹문서(상품)를 검색하고자 시도할 경우 협력적 여과 시스템에 웹문서가 존재하지 않을 지라도 새로운 사용자에게 웹문서를 추천할 수 있는 장점을 갖는다.

본 논문에서는 내용 기반 여과를 위하여 Naive Bayes 분류자에 의한 문서 분류 방법에 따라 클래스별로 분류된 웹문서를 저장한 데이타베이스를 사용한다. 데이타베이스의 색인은 연관 단어 마이닝의 방법에 따라 웹문서로부터 추출된 특징이다. 내용 기반 여과에서의 검색 방식은 사용자 중심의 지능형 점보검색 시스템[12]의 검색 방법을 사용한다.

2.1 문서의 특징 표현

본 논문에서는 문서의 특징을 표현하기 위해 '단일 단어가 든 가방[13]'의 형태를 이용한 '연관 단어가 든 가방(bag-of-association words)'의 형태를 사용한다[14, 15]. 이러한 형태로 문서의 특징을 표현하기 위해 형태소 분석을 통한 명사 추출 과정을 전처리 과정으로 사용한다. 형태소 분석에 사용되는 시스템은 사용자 중심의 지능형 점보검색 시스템[12]에서 사용된 방법이다. Apriori 알고리즘[16, 17]은 형태소 분석에 의해 추출된 명사들로부터 연관 단어를 마이닝한다. 연관 규칙 마이닝 알고리즘인 Apriori는 구매하는 물품들의 집합인 트랜잭션으로부터 연관 규칙을 마이닝한다. 마이닝한 결과, 각 문서는 연관 단어들의 집합, 즉 연관 단어 벡터 모델로 나타내어진다. 이에 따라 문서$\{d_j\}$는 식 (1)과 같이 연관 단어 벡터 모델로 표현된다.

$\{d_j\}=\{(w_{j1}\&w_{j2}...\&w_{j(r-1)} \Rightarrow w_{jr}),(w_{j1}\&w_{j2}...\&w_{j(r-1)} \Rightarrow w_{jr}),...,(w_{k1}\&w_{k2}...\&w_{k(r-1)} \Rightarrow w_{kr}),...,(w_{p1}\&w_{p2}...\&w_{p(p-1)} \Rightarrow w_{pr})\}$, $(j=1,2,...,m)$ (1)

식 (1)에서 $(w_{j1}\&w_{j2}\&w_{j(r-1)}\&w_{jr})$는 연관 단어, $\{w_{j1},w_{j},w_{j(r-1)},w_{jr}\}$는 연관 단어를 구성하는 단어들의 구성, r은 연관 단어를 구성하는 단어의 수, m은 텍스트를 대표하는 연관 단어의 수이다. 또한 $\&$는 각 단어와 단어가 연관 되었음을 의미하는 기호이다. 연관 단어를 가장 적절하게 추출하기 위해서 신뢰도는 85보다 크도록 지지도는 22보다 작도록 지정해야 한다[10].

2.2 Nave Bayes 분류자를 사용한 분류

Naive Bayes 분류자는 학습 단계와 분류 단계를 통하여 문서를 분류한다[18, 19]. 학습 단계에서는 Apriori 알고리즘에 의해 구축된 연관 단어 지식베이스[20]의 연관 단어에 추정치를 부여한다. 식 (2)는 연관 단어 지식베이스의 $classID$에 있는 k번째 연관 단어$(w_{k1}\&w_{k2}...\&w_{k(r-1)}\Rightarrow w_{kr})$에게 추정치를 부여하기 위한 식이다.

$$P((w_k\&w_{k2}...\&w_{k(r-1)} \mid classID)=\frac{n_{k+1}}{n+|AWKB|} \quad (2)$$

식 (2)에서 $P((w_{k1}\&w_{k2}\&w_{k(r-1)}\Rightarrow w_{kr})|classID)$는 $classID$에 있는 k번째 연관 단어 $(w_{k1}\&w_{k2}...\&w_{k(r-1)}\Rightarrow w_{kr})$의 추정치, n은 학습문서로부터 마이닝된 연관 단어의 전체 수, n_k는 전체 개수 n 중에서 지식베이스에 있는 연관 단어 $(w_{k1}\&w_{k2}\&w_{k(r-1)}\Rightarrow w_{kr})$와 일치하는 연관 단어의 수이다. 또한, $classID$는 연관 단어 지식베이스에 있는 클래스의 레이블이며, $|AWKB|$는 연관 단어 지식베이스에 있는 전체 연관 단어의 수이다.

학습 과정은 누적 단계와 추정치 부여 단계로 나눈다. 누적 단계에서는 학습문서에 있는 연관 단어가 지식베이스 안에 있는 경우 횟수를 누적시킨다. 추정치 부여 단계에서는 누적 단계의 결과를 식 (2)에 적용하여 지식베이스의 연관 단어에 추정치를 부여한다. 이러한 과정을 통해, 지식베이스의 연관 단어에 추정치가 추가된다. 표 2는 훈련문서로 추출된 연관 단어에 식 (2)를 사용하여 추정치를 추가한 결과이다.

분류 단계에서는 추정치가 부여된 연관 단어 지식베이스를 사용하여 Naive Bayes 분류자에 의해 실험문서를 클래스로 분류할 수 있다. Apriori 알고리즘은 2.1절에서 기술한 방법으로 실험문서로부터 연관 단어의 형태로 특징을 추출한다.

실험문서(D)의 특징이 $\{d(w_{j1}\&w_{j2}...\&w_{j(r-1)}\Rightarrow w_{jr}).d(w_{j1}\&w_{j2}...\&w_{j(r-1)}\Rightarrow w_{jr}),...,d(w_{k1}\&w_{k2}...\&w_{k(r-1)}\Rightarrow w_{kr}),...,d(w_{m1}\&w_{m2}...\&w_{m(r-1)}\Rightarrow w_{mr})\}$라고 하였을 때 Naive Bayes 분류자는 식 (3)을 이용하여 실험문서를 클래스로 분류한다.

표 2 게임 클래스의 연관 단어에 추정치 부여

연관 단어	추정치
게임& 구성& 선수& 경기& 스포츠& 참가=> 선발	0.093750
국내& 최신& 기술& 설치-> 개발	0.012700
게임& 참가& 인기& 사용자& 접속=> 이벤트	0.100386
운영& 선발& 경기& 순위& 규칙=> 평가	0.016878
게임& 순위& 이름=> 스포츠	0.089494
운영& 스포츠& 위원회& 선수=> 선발	0.100386
게임& 구성& 선발& 순위=> 경기	0.086614
게임& 일정& 선수& 참가& 운영=> 스포츠	0.093023

$d(w_{k1}\&w_{k2}...\&w_{k(r-1)}\Rightarrow w_{kr})$의 "$d$"는 실험문서로부터 추출된 연관 단어임을 나타낸다.

$$class=\arg\max_{classID=1}^{N} P(classID)\prod_{k=1}^{m} P(d(w_{k1}\&w_{k2}\&w_{k(r-1)}\Rightarrow w_{kr})|classID) \quad (3)$$

식 (3)에서 전체 클래스의 수는 N이며 그 중 문서D가 분류될 클래스는 $class$이다. 또한 $P(d(w_{k1}\&w_{k2}...\&w_{k(r-1)}\Rightarrow w_{kr})|classID)$는 식 (2)에 의해 계산된 추정치이며, $P(classID)$는 $classID$로 분류될 확률이다.

2.3 사용자 프로파일 생성

내용 기반 여과 시스템은 사용자에게 추천된 웹문서 중에서 연관 피드백 방법에 의해 사용자가 방문한 웹문서를 대상으로 새로운 사용자의 프로파일을 만든다[21]. 연관 피드백 방법에 의해 사용자가 방문한 웹문서로부터 추출된 연관 단어는 사용자의 프로파일에 누적된다. 사용자의 프로파일에 누적된 연관 단어의 발생 빈도에 따라 프로파일에 있는 연관 단어에 가중치가 부여된다. 가중치가 부여된 새로운 사용자의 프로파일을 WU_a로 정의하였을 경우 WU_a는 식 (4)와 같이 표현된다.

$$WU_a=\{w_1AW_1,w_2AW_2,...,w_tAW_t\} \quad (4)$$

식 (4)에서 $\{w_1,w_2,...,w_t\}$는 각 연관 단어의 가중치를 나타내는 가중치 벡터이며, t는 사용자 프로파일에 있는 연관 단어의 모든 종류의 수이다.

3. 협력적 여과

협력적 여과에서는 Naive Bayes 알고리즘과 유전자 알고리즘을 이용하여 사용자 행렬의 차원수를 줄인다. 기존의 군집화 알고리즘은 사전에 군집의 개수를 미리 결정해야 하고 잡음에 민감하며 지역적 최적해에 수렴할 수 있다는 문제점을 가지고 있다[22]. 이러한 문제점의 개선을 위해 본 논문에서는 유전자 알고리즘을 이용

한 군집화 기법을 사용한다.

3.1 상품 분류

본 논문에서는 협력적 여과 추천에서의 {사용자-상품}의 행렬을 $R=(r_{ij})(i=1,2,...,n \text{ and } j=1,2,...,m)$로 표현한다. 행렬 R은 m개의 상품(item)과 n명의 사용자(user)의 집합으로 구성되며, r_{ij}는 행렬의 값으로 사용자 u_i의 상품 d_j에 대한 선호도를 나타낸다. 구체적으로, 상품 집합은 $I=\{d_j\}(j=1,2,...,m)$로 기술하며 사용자 집합은 $U=\{u_i\}(i=1,2,...,n)$로 표현한다. 사용자의 선호도를 발견하기 위하여 사용자가 평가한 상품 중에서 사용자에게 흥미롭다고 평가된 상품만을 분류의 대상으로 한다. 평가는 0, 0.2, 0.4, 0.6, 0.8 그리고 1의 6단계로 이루어졌으며 0.5보다 크게 평가되었을 경우 사용자가 흥미롭다고 분류하였다. 사용된 상품은 웹문서 수집기에 의해 수집된 컴퓨터 분야의 웹문서이다. 웹문서의 특징은 2.1절에서 기술된 연관 단어 마이닝 방법을 사용하며 상품의 분류는 2.2절에서 설명된 Naive Bayes 알고리즘을 사용한다.

3.2 최적의 사용자 군집

본 장에서는 3.1장에서 분류된 상품을 기반으로 유전자 알고리즘을 사용하여 사용자를 군집하는 방법을 제시한다. 이를 위하여 3.1절에서 정의된 행렬 R를 행렬 R_1으로 변환한다. 행렬 $R_1=(c_{ki})(k=1,2,...,N \text{ and } i=1,2,...,n)$는 {클래스-사용자}의 행렬로 R_1에서 클래스 $C=\{class_k\}(k=1,2,...,N)$는 각 클래스로 분류된 상품의 집합이라고 정의한다. 행렬 R_1의 값 c_{ki}는 사용자 u_i가 흥미롭다고 평가한 상품 중에서 $class_k$로 분류된 상품의 수이다. 이러한 정의를 기반으로 행렬 R_1을 구성하기 위하여 식 (5)를 정의한다. 식 (5)는 3.1절에서 정의한 행렬 R의 값 r_{ij}가 0.5이상인 상품 $d_j(0<j\leq m)$가 클래스 $class_k(0<k\leq N)$에 속할 때 행렬 R_1의 값 c_{ki}의 값을 하나씩 누적시키는 식이다.

$$R_1 = \{d_j \in class_k, 0<j\leq m, 0<class_k\leq N \mid C_{ki} = C_{ki}+1, \forall r_{ij}>0.5\} \quad (5)$$

표 3은 18명의 사용자가 흥미롭다고 평가한 상품을 Naive Bayes 분류자에 의해 10개의 클래스로 분류한 결과에 대해 식 (5)를 적용함으로써 생성된 행렬 R_1의 예를 보인다.

행렬 R_1의 구조를 기반으로 유전자 알고리즘은 사용자들을 군집하여 그룹을 생성한다. 유전자 알고리즘은 유전인자(Gene), 염색체(Chromosome), 모집단(Population)을 사용함으로써 사용자를 군집한다. 모집단은 초기화, 적합도 계산, 재구성, 선택, 교배, 돌연변이 그리고 평가의 과정을 거쳐 진화한다.

표 3 행렬 R_1의 예

클래스	사용자																	
	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}
1	3	0	1	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0
2	0	0	2	0	0	0	0	0	0	0	0	2	0	0	1	0	0	1
3	2	0	4	0	1	1	1	2	0	2	0	0	0	0	2	2	1	3
4	2	2	5	4	2	2	3	1	0	1	0	0	0	0	1	1	2	1
5	1	0	1	0	2	0	1	3	0	3	0	0	0	0	3	2	1	2
6	2	3	1	5	3	3	2	1	0	3	0	0	0	0	1	4	3	1
7	3	0	0	0	1	1	1	2	0	0	0	0	0	0	0	0	0	0
8	1	4	1	1	1	1	1	4	0	1	0	0	0	0	1	1	1	1
9	1	1	2	1	2	0	2	1	0	1	0	1	3	3	3	2	2	2
10	1	0	1	0	1	2	0	1	1	0	0	0	0	0	0	0	1	0

초기화는 각 개체를 표현하는 단계이다. 유전자 알고리즘에서는 문제의 가능한 해를 직접적으로 탐색하지 않고 염색체의 형태로 바꾸는 인코딩 과정을 거쳐 코딩된 개체에 대해서 진화 연산자를 적용시킴으로써 탐색을 수행한다. 인코딩에 의해서 해를 나타내므로 주어진 문제를 어떤 식으로 코딩하느냐에 따라서 해공간의 형태가 바뀌게 되고, 적용할 수 있는 진화 연산자의 종류가 달라지게 되며, 또한 진화 연산자에 의해서 만들어진 탐색 영역도 달라지게 된다. 따라서 효율적인 탐색을 위해서는 개체의 표현 방법이 매우 중요하다. 유전자 알고리즘에서 사용되는 염색체의 표현 방법은 이진 문자열, 실수 벡터, 트리 형태 등이 있다.

본 논문에서는 표 3과 같은 행렬을 염색체로 표현해야 하므로 이진 문자열, 실수 벡터, 트리 형태 중 실수로 표현하는 것이 가장 적합하다. 실수로 표현된 염색체는 최적의 해를 찾는 과정에서 각 염색체의 적합율을 계산하게 된다. 계산 과정에서의 복잡도를 줄이고 정확도를 높이기 위하여 행렬 R_1의 값 c_{ki}를 식 (6)에 의해 c'_{ki}로 변환시킨다.

$$c'_{ki} = log_2(\sqrt{c_{ki}}+1) \quad (6)$$

식 (6)은 c_{ki}의 값을 줄이기 위하여 c_{ki}의 값에 로그와 제곱근을 사용하며 c_{ki}의 값이 0일 경우를 방지하기 위하여 1를 더한다.

행렬 R_1의 값 c_{ki}를 식 (6)에 의해 c'_{ki}로 변환함에 따라 행렬 R_1은 행렬 R'_1으로 재정의된다. 이에 따라 행렬 R_1'은 초기화 과정을 거쳐 생성된 개체이다. 표 4는 표 3의 행렬 R_1을 식 (6)을 사용하여 행렬 R_1'으로 변환한 결과이며 또한 초기화에 의해 생성된 개체이기도 하다. 유전인자는 행렬 R_1'에 있는 요소 c'_{ki}을, 염색체는 행렬 R_1'에 있는 각 줄을 나타낸다. 그리고 모집단은

표 4 행렬 R_1'의 요소

클래스	염색체에 속한 각 유전인자																	
	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}
염색체1	1.45	0	1	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0
염색체2	0	0	1.27	0	0	0	0	0	0	0	0	1.27	0	0	1	0	0	1
염색체3	1.27	0	1.58	0	1	1	1	1.27	0	1.27	0	0	0	0	1.27	1 27	1	1.45
염색체4	1.27	1.27	1.69	1.58	1.27	1.27	1.45	1	0	1	0	0	0	0	1	1	1.27	1
염색체5	1	0	1	0	1.27	0	1	1.45	0	1.45	0	0	0	0	1.45	1.27	1	1.27
염색체6	1.27	1.45	1	1.69	1.45	1.45	1.27	1	0	1.45	0	0	0	0	1	1.58	1.45	1
염색체7	1.45	0	0	0	1	1	1	1.27	0	0	0	0	0	0	0	0	0	0
염색체8	1	1.58	1	1	1	1	1	1.58	0	1	0	0	0	0	1	1	1	1
염색체9	1	1	1.27	1	1.27	0	1.27	1	0	1	0	1	1.45	1.45	1.45	1.27	1.27	1.27
염색체10	1	0	1	0	1	1.27	0	1	1	0	0	0	0	0	0	0	1	0

행렬 R_1'에 있는 모든 요소를 나타낸다.

유전자 알고리즘에서는 개체의 성능을 다른 개체와 비교하기 위하여 적합율을 모든 개체에 부여한다. 본 논문에서는 클래스에 분류된 아이템의 분포가 비슷한 클래스를 찾아내고 사용자들의 최적화하여 군집하기 위해 적합율을 계산한다. 사용자를 군집하기 위해서는 상호간의 유사도가 높은 사용자를 찾아야 한다. 이를 위해 사용자간의 유사도[23]는 정보 검색 분야에서 문서간의 유사도를 구하기 위해 사용되는 식 (7)의 Jaccard 방법[9]을 이용한다. 적합율을 계산하기 위한 식 (7)에서 c'_{ki}는 행렬 R'_1의 각 요소를 나타내며 c''_{ki}는 행렬 R_1'에서 c'_{ki}를 제외한 모든 요소를 나타낸다.

$$Fitness(c'_{ki}) = \left(\sum_{i=1}^{N} \frac{\sum_{k=1}^{N} c'_{ki} c''_{ki}}{\sum_{k=1}^{N} c'^{2}_{ki} + \sum_{k=1}^{N} c''^{2}_{ki} - \sum_{k=1}^{N} c'_{ki} c''_{ki}} \right) / N \quad (7)$$

재구성 단계에서는 적합율 계산을 위한 목적 함수의 값을 다른 값으로 재구성한다. 재구성하는 목적은 적합율을 선택연산자를 적용하는 확률로 사용하기 위함이다. 본 논문에서는 적합율의 조정을 위해 식 (8)을 이용한다. 식 (8)은 전체 염색체의 적합율의 합에 대한 각 염색체의 적합율의 비를 계산한다. 식 (8)에서 $Fitness(c'_{ki})$는 각 염색체의 적합율이며 $Fitness_s(c'_{ki})$는 염색체의 적합율을 재구성한 값이다.

$$Fitness\_s(\boldsymbol{C}'_{ki}) = \frac{Fitness(\boldsymbol{C}'_{ki})}{\sum_{class=1}^{N} Fitness(\boldsymbol{C}'_{ik})} \quad (8)$$

표 5는 표 4에 나타난 10개 염색체의 적합율과 재구성 적합율을 나타낸다. 염색체의 적합율은 식 (7)에 의해 계산되며, 재구성 적합율은 식 (8)에 의해 계산된다.

표 5 염색체의 적합율 및 재구성 적합율

염색체	적합율	재구성 적합율
1	0.267916	5.616920
2	0.251901	5.281176
3	0.604733	12.67838
4	0.57429	12.04014
5	0.563645	11.81695
6	0.58829	12.33365
7	0.375081	7.863669
8	0.612715	12.84572
9	0.526596	11.04021
10	0.404631	8.483185
평균 적합율	0.517669	10

선택 단계에서는 선택 연산자가 재구성된 적합율을 기반으로 교배가 될 염색체를 선택한다. 표 5에서 염색체3, 염색체4, 염색체5, 염색체6, 염색체8, 염색체9는 재구성 적합율이 다른 염색체보다 크므로 나머지 염색체보다 선택될 확률이 높다.

교배 단계에서는 교배 연산자가 교배율을 기반으로 선택 연산자에 의해 선택된 염색체를 대상으로 교배를 시킨다. 교배 방식은 임의의 한 유전인자를 선택하여 선택된 유전인자가 위치한 지점부터 마지막까지의 유전인자까지 유전인자의 값을 교환하는 1점 교배 방식이다. 통상적으로 교배율(Pc)은 0.7-0.9의 값을 사용하며 본 논문에서 사용된 교배율은 0.9이다.

돌연변이 단계에서는 하나의 비트를 주어진 확률에 따라서 다른 값으로 바꾸는 단계이다. 돌연변이율은 염색체의 유전인자가 다른 값으로 바뀌는 확률을 의미한다. 돌연변이율은 0.01-0.05사이의 값을 많이 사용하며 본 논문에서는 돌연변이율을 0.01로 정한다.

표 6은 표 4의 부모 염색체가 선택, 교배, 돌연변이 단계를 거쳐 탄생된 2세대의 염색체를 나타낸다.

표 6 제 2세대 염색체 및 적합율

	염색체에 있는 유전인자(2세대)																	적합율
1	1.27	0	1.58	0	1	1	1	1.27	0	1.27	0	0	0	0	1.27	1.27	1	1.45	0.73157825
2	1.27	1.27	1.69	1.58	1.27	1.27	1.45	1	0	1	0	0	0	0	1	1	1.27	1	0.74318470
3	1.27	0	1.58	0	1	1	1	1.27	0	1.27	0	0	0	0	1.27	1.27	1	1.45	0.73157825
4	1	0	1	0	1.27	0	1	1.45	0	1.45	0	0	0	0	1.45	1.27	1	1.27	0.68194970
5	1.27	1.45	1	1.69	1.45	1.45	1.27	1	0	1.45	0	0	0	0	1	1.58	1.45	1	0.72893251
6	1.45	0	0	0	1	1	1	1.27	0	0	0	0	0	0	0	0	0	0	0.42592328
7	1	1.58	1	1	1	1	1	1.58	0	1	0	0	0	0	1	1	1	1	0.74183814
8	1	1	1.27	1	1.27	0	1.27	1	0	1	0	1	1.45	1.45	1.45	1.27	1.27	1.27	0.60842548
9	1.27	1.45	1	1.69	1.45	1.45	1.27	1	0	1.45	0	0	0	0	1	1.58	1.45	1	0.72893251
10	1	0	1	0	1	1.27	0	1	1	0	0	0	0	0	0	0	1	0	0.45097332
	평균 적합율 = 0.47698																		

표 7 10세대에서 120세대까지의 염색체와 각 세대의 평균 적합률

진화 세대	염색체의 유전인자																	재구성 적합률	평균
10	1 27	0	1.58	0	1	1	1	1.27	0	1.27	0	0	0	0	1.27	1.27	1	1.45	11.6589	
	1 27	1.27	1.69	1.58	1.27	1.27	1.45	1	0	1	0	0	0	0	1	1	1.27	1	11.2508	0.6573
									. .										. .	
20	1.27	0	1.58	0	1	1	1	1.27	0	1.27	0	0	0	0	1.27	1.27	1	1.45	11.4441	0.671
	1.27	1 27	1.69	1.58	1.27	1.27	1.45	1	0	1	0	0	0	0	1	1	1.27	1	10.6296	
									. .										. .	
30	1.27	0	1.58	0	1	1	1	1.27	0	1.27	0	0	0	0	1.27	1.27	1	1.45	10.9952	
	1.27	0	1.58	0	1	1	1	1.27	0	1.27	0	0	0	0	1.27	1.27	1	1.45	10.9952	0.6877
																			. .	
40	1.27	0	1.58	0	1	1	1	1.27	0	1.27	0	0	0	0	1.27	1.27	1	1.45	10.9952	
	1.27	1.27	1.69	1.58	1.27	1.27	1.45	1	0	1	0	0	0	0	1	1	1.27	1	9.97998	0.7332
									. .										. .	
50	1	0	1	0	1.27	0	1	1.45	0	1.45	0	0	0	0	1.45	1.27	1	1.27	10.9666	
	1.27	0	1.58	0	1	1	1	1.27	0	1.27	0	0	0	0	1.27	1.27	1	1.45	11.2051	0.7632
									. .										. .	
60	1	0	1	0	1.27	0	1	1.45	0	1.45	0	0	0	0	1.45	1.27	0	1.27	10.7554	
	1.27	0	1.58	0	1	1	1	1.27	0	1.27	0	0	0	0	1.27	1.27	1	1.45	11.2535	0.8027
									. .										. .	
70	1.27	0	1.58	0	0	0	1	1.27	0	1.27	0	0	0	0	1.27	1.27	0	1.45	12.7772	
	1.27	0	1.58	0	0	1	0	1.27	0	1.27	0	0	0	0	1.27	1.27	0	1.45	12.7772	0 8381
									. .										. .	
80	1 27	1.45	1	1.69	1.45	1.45	1.27	1	0	1.45	0	0	0	0	1	1.58	1.45	1	10.0512	
	1.27	0	1.58	0	0	0	1	1.27	0	1.27	0	0	0	0	1.27	1.27	0	1.45	13.0385	0.911
									. .										. .	
90	1.27	0	1.58	0	1	0	0	1.27	0	1.27	0	0	0	0	1.27	1.27	0	1.45	13.4169	
	1.27	0	1.58	0	0	0	0	1.27	0	1.27	0	0	0	0	1.27	1.27	0	1.45	13.4169	0.9826
									. .										. .	
100	1.27	0	1.58	0	0	0	0	1.27	0	1.27	0	0	0	0	1.27	1.27	0	1.45	1	
	1.27	0	1.58	0	0	0	0	1.27	0	1.27	0	0	0	0	1.27	1.27	0	1.45	1	1
									. .										1	
	1.27	0	1.58	0	0	0	0	1.27	0	1.27	0	0	0	0	1.27	1.27	0	1.45	1	

평가 단계에서는 진화를 계속할 것인가 종료할 것인가가 결정되는 단계이다. 평가하는 기준은 평균 적합율이 적합율 임계값과 같거나 크다면 진화를 종료하고 작다면 재구성 단계부터 진화를 반복한다. 본 논문에서는 적합율 임계값을 1로 하여 계산된 평균 적합율이 1보다 작다면 진화를 계속 진행하고 1보다 크거나 같다면 종료한다. 표 6에서 3열에 계산된 2세대 염색체의 평균 적합율은 0.47698이므로 평가 기준에 따라 다음 세대로의 진화가 계속된다.

표 7은 10세대에서 마지막 세대인 100세대까지의 10세대간격의 염색체와 세대의 평균 적합율을 보인다.

진화가 진행되는 동안 표 7의 3열에서와 같이 평균 적합율은 일정하게 증가되어 평균 적합율이 1이 되는 100세대에서 진화는 종료되었다. 그림 1은 1세대에서 100세대까지의 평균 적합률의 변화를 나타낸다.

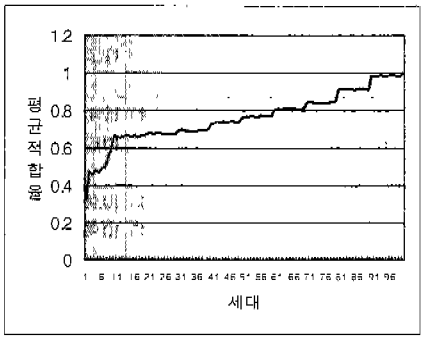

그림 1 세대별 평균 적합율의 변화

그림 2는 1세대에서 100세대까지의 세대별 수행 시간을 나타낸다. 본 논문에서는 유전자 알고리즘을 2-way CPU와 256MB의 주기억 장치, 20GB의 보조 기억장치로 구성되어 있는 삼보 SMP NT서버 6400Qp상에서

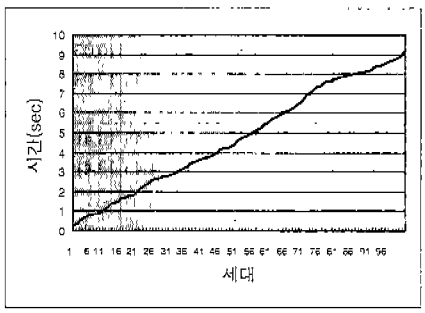

그림 2 세대별 수행 시간

구현하였다. 이러한 구현 환경에서 1세대에서 100세대까지의 최적의 사용자 군집에 걸린 시간은 9.13초이다.

100세대의 염색체에서 1이상의 유전인자를 나타내는 사용자는 채택되고 0의 유전인자를 나타내는 사용자는 군집의 대상에서 제거된다. 이에 따라 첫번째 군집은 7명의 사용자로 구성된다. 첫번째 그룹으로 군집된 사용자를 제외하고 나머지 사용자를 대상으로 초기화, 적합율 계산, 재구성, 선택, 교배, 돌연변이, 평가의 과정을 반복한다. 이러한 반복 과정을 통하여 사용자를 각각의 그룹으로 군집시킬 수 있다. 결과적으로, 유전자 알고리즘은 표 4에 나타난 18명의 사용자를 표 8과 같이 4개의 그룹으로 군집시킨다.

표 8 군집된 사용자 그룹

그룹	사용자
1	user1, user3, user8, user10, user15, user16, user18
2	user2, user4, user5, user17
3	user11, user12, user13, user14
4	user6, user7, user9

3.3 그룹 프로파일의 생성

본 논문에서의 협력적 여과 추천 시스템에서는 새로운 사용자와 가장 유사한 그룹을 검색하고, 검색된 그룹을 대상으로 그들이 구입한 상품목록에서 사용자에게 추천할 상품 목록을 선정한다. 새로운 사용자와 각 그룹의 유사도는 사용자 프로파일과 그룹 프로파일을 각각 생성한 후 프로파일간의 유사도를 계산함으로써 얻는다. 그룹 프로파일은 그룹에 속한 사용자들이 소유한 특징을 추출하여 누적시킴으로써 생성한다. 그룹 내의 사용자들의 특징은 사용자들이 흥미롭다고 평가한 상품으로부터 추출된 연관 단어의 집합이다. 이러한 경우 상품으로부터의 특징을 추출하기 위하여 2.1절에 나타난 연관 단어 마이닝 방법을 사용한다. 사용자가 군집된 그룹 중 하나를 그룹 $\ell(C')$이라고 할 경우 그룹 ℓ은 식 (9)로 표현한다. 식 (9)에서 그룹 ℓ은 연관 단어로 구성된 특징 벡터이다.

$$C^{\ell} = \{AW_1^{\ell}, AW_2^{\ell}, \ldots, AW_r^{\ell}\} \quad (9)$$

식 (9)에서 그룹 ℓ에 속한 연관 단어의 수는 r이며, AW_r^{ℓ}는 그룹 ℓ에 속한 연관 단어라고 정의한다.

그룹 프로파일에 속한 연관 단어는 사용자가 흥미롭다고 평가한 상품으로부터 추출된 연관 단어의 빈도에 따라 가중치를 부여한다. 이에 따라 식 (9)으로 표현된 그룹 ℓ의 각 연관 단어에 가중치를 부여할 경우 가중치가

부여된 그룹 $\ell(WC\ ')$은 식 (10)에 의해 정의한다.

$$WC^l = \{k_1^l AW_1^l, k_2^l AW_2^l, ..., k_r^l AW_r^l\} \quad (10)$$

여기서, $\{k_1^l, k_2^l, ..., k_r^l\}$는 모든 그룹내의 연관 단어의 총 수에 대한 각 연관 단어의 수의 비율을 나타내는 가중치 벡터이다. 표 9는 표 8에 속한 사용자로부터 추출된 가중치가 부여된 그룹 프로파일의 일부를 보인다.

표 9 가중치가 부여된 그룹 프로파일

그룹	그룹 프로파일
1	게임&구성&선수&경기&스포츠&참가=>선발(0.231) 국내&최신&기술&설치=>개발(0.3291) ..
2	방법&중심&제작=>사용(0.1938) 뉴스&제공&홍=>속보(0.8292)..
3	시스템&사업&활용=>기법(0.382) 세계&네티즌&인물&조작=>수법(0.281)
4	입력&편집&출력&컬러&종류(0.2131) 보드&주변기기&슬롯&펜티엄(0.321)

4. 사용자 선호도 발견

그림 3은 협력적 추천 시스템과 내용 기반 여과를 병합하여 사용자의 선호도를 발견하는 시스템이다.

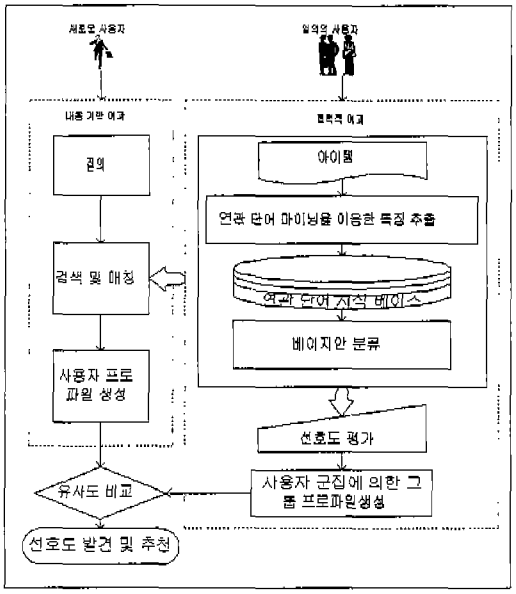

그림 3 협력적 추천 시스템과 내용 기반 여과를 병합한 사용자의 선호도 발견

그림 3에서 내용 기반 여과 추천은 2장에서 설명 부분으로 사용자의 질의가 입력되면 질의를 기반으로 상품을 검색하며 검색된 결과를 바탕으로 연관 피드백을 사용함으로써 사용자의 프로파일을 생성한다.

협력적 여과 추천은 3장에서 기술된 내용으로 먼저 상품을 베이지안 분류에 의해 클래스로 분류함으로써 상품의 차원수를 줄이며, 이를 기반으로 유전자 알고리즘을 사용하여 사용자를 군집시킴으로써 사용자에 대한 차원수를 감소시킨다.

새로운 사용자의 선호도는 내용 기반 여과 추천에서 생성된 사용자 프로파일과 협력적 여과 추천에서 생성된 그룹 프로파일을 병합함으로써 발견할 수 있다. 새로운 사용자의 프로파일과 그룹 프로파일을 병합하기 위해서는 우선 새로운 사용자가 협력적 여과 추천에서 생성된 각 그룹 중 하나의 그룹으로 분류되어야 한다. 이를 위하여 본 논문은 그룹 프로파일과 새로운 사용자의 프로파일과의 유사도를 계산하여 가장 유사도가 큰 그룹으로 사용자를 분류한다. 그룹 프로파일과 새로운 사용자의 유사도를 계산하기 위해 본 논문에서는 정보검색 분야에서 사용되는 벡터유사도를 사용한다[25, 26]. 식 (11)은 가중치가 부여된 새로운 사용자의 프로파일 WU_a와 각 그룹 프로파일 WC^l간의 벡터유사도를 계산하는 식이다. 식 (11)에서 그룹 프로파일에 있는 연관 단어가 사용자 프로파일에 없다면 연관 단어의 가중치는 0이며, 있을 경우 가중치는 이 된다.

$$Similarity(WC^l, WU_a) = \frac{\sum_{i=1}^{r} k_i^l w_r}{\sqrt{\sum_{i=1}^{r} k_i^{l\,2}} \sqrt{\sum_{i=1}^{r} w_r^2}} \quad (11)$$

식 (11)을 이용하여 새로운 사용자의 프로파일과 각 그룹 프로파일간의 벡터유사도를 계산하였을 경우 새로운 사용자는 가장 큰 유사도를 갖는 그룹으로 분류된다.

내용 기반 여과 시스템에서 추천된 문서 중에서 새로운 사용자가 방문한 2개의 웹문서가 있다고 가정할 경우 표 10은 이들 2개의 문서에 대한 특징과 Naive Baycs 분류자에 의해 분류된 클래스를 보인다.

표 10 웹문서의 특징 및 분류된 클래스

웹문서	특 징	클래스
1	국내&최신&기술&설치=>개발 게임&참가&인기&사용자&전속=>이벤트 운영&선발&경기&순위&규칙=>평가 게임&순위&이름=>스포츠 운영&스포츠&위원회&선수=>선발	1
2	뉴스&증진=>비전 발표&라디오=>통신 텔레비전&뉴스=>신문	3

표 10을 기반으로 새로운 사용자의 프로파일은 다음과 같이 구성한다.

Profile={국내&최신&기술&설치=>개발,게임&참가&인기&사용자&접속=>이벤트,운영&선발&경기&순위&규칙=>평가,게임&순위&이름=>스포츠, 운영&스포츠&위원회&선수=>선발, 뉴스&증진=>비젼,발표&라디오=>통신,텔레비전&뉴스=>신문}.

표 10의 새로운 사용자는 식 (11)에 의한 계산 결과 그룹 1로 분류되었다. 분류된 결과를 기반으로 새로운 사용자는 사용자가 평가하지 않은 상품을 추천 받을 수 있다. 추천 받을 상품은 그룹 1에 속한 사용자가 흥미롭다고 평가한 상품 중에서 class1과 class3으로 분류된 상품이다. 표 8은 그룹 1에 7명의 사용자를 보이며 표 3은 class 1과 class 3에 총 15개의 상품을 보인다. 상품의 빈도를 계산하여 추천을 위한 사용자 선호도의 가중치로 사용한다. 각 상품은 가중치의 내림차순으로 정렬되며 정렬된 상품은 새로운 사용자에게 추천된다.

5. 성능 평가

협력적 여과 추천을 위한 데이타베이스는 200명의 사용자와 1600개의 서로 다른 웹문서로 구성한다. 사용자는 1600개의 웹문서에 대해 적어도 10개의 평가를 한 사용자들이다. 내용 여과 기반의 추천을 위한 데이타베이스는 1600개의 서로 다른 웹문서로 구성한다. 1600개의 웹문서는 웹문서수집기에 의해서 컴퓨터분야의 URL로부터 수집된 문서이다. 1600개의 훈련문서는 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. 200명의 사용자 중 100명의 사용자는 훈련을 위해 사용되며 나머지 100명은 테스트를 위해 사용된다.

추천의 성능을 평가하기 위해 본 논문에서는 [27]에 의해 제안된 MAE(Mean Absolute Error)와 순위 스코어 측정(Rank scoring metric)을 사용한다. MAE는 단일 상품의 추천 시스템을 평가하는 데 사용하며, 순위 스코어 측정은 순위가 있는 상품의 목록을 추천하는 시스템의 성능을 평가하는 데 사용한다. MAE에서 예측의 정확도는 실제로 사용자가 평가한 값과 예측된 값의 차이에 대한 절대값의 평균을 나타내며 식 (12)에 의해 정의된다.

$$s_a = \frac{1}{m_a}\sum_{i\in} p_a \left| p_{a,j} - v_{a,i} \right| \qquad (12)$$

식 (12)에서 $p_{a,i}$는 예측된 선호도이며 $v_{a,i}$는 실제로 사용자가 평가한 선호도이다. 또한 m_a는 새로운 사용자에 의해 평가된 상품의 수를 의미한다.

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

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

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

$$R = 100 \times \frac{\sum_a R_a}{\sum_a R_a^{\max}} \qquad (14)$$

식 (14)에서 $R_a^{\max}$는 사용자가 평가하거나 방문한 상품이 순위가 있는 목록상에서 상위에 나타났을 경우에 측정된 순위 스코어 측정에 대한 기대 이용도의 최대값이다.

본 논문에서는 평가를 위해 제안된 협력적 여과와 내용 기반 여과를 병합한 추천 방법(C_G_N), 피어슨 상관계수를 이용한 기존의 메모리 기반 방법(P_Corr), 내용 기반 여과만을 이용한 추천 방법(Content), 본 논문에서 제안된 방법 중 협력적 여과만을 이용한 추천 방법(G_N)을 군집된 사용자의 수를 변화시키면서 성능을 비교하였다. 또한 제안한 방법을 표 1의 협력적 여과와 내용 기반 여과를 병합한 기존 추천 시스템 중 희박성을 해결한 Soboroff의 방법, 희박성과 초기 평가 문제를 해결한 Pazzani 방법, 초기 평가 문제를 해결한 Fab 방법과 사용자가 상품에 대해 평가한 횟수를 변화시켜가면서 비교하였다.

표 11은 식 (12)와 식 (14)를 기반으로 사용자 수를 변화시킴에 따른 C_G_N, P_Cor, Content, G_N의 MAE와 순위 스코어를 나타낸다.

그림 4과 그림 5는 표 11를 기반으로 한 사용자의 수에 따른 MAE와 순위 측정 척도를 나타낸다. 그림 4와

표 11 사용자 수의 변화에 따른 MAE와 순위 스코어

사용자 수	MAE				순위 스코어			
	P_Corr	G_N	Content	C_G_N	P-Corr	G_N	Content	C_G_N
10	0.210	0.210	0.230	0.190	56	52	55	56
20	0.210	0.206	0.221	0.180	56.1	53	54	57
30	0.208	0.200	0.221	0.183	56.2	55	54	58
40	0.207	0.198	0.219	0.182	56.8	56	55	59
50	0.204	0.191	0.219	0.181	57	57	54	61
60	0.203	0.188	0.218	0.178	57.1	58	55	62
70	0.202	0.185	0.218	0.175	58	58.5	55	64
80	0.201	0.180	0.218	0.170	58.1	59	53	65
90	0.201	0.175	0.218	0.165	58.1	59.5	54	65.5
100	0.200	0.170	0.218	0.160	58.1	60	54	66

그림 5는 사용자들의 수가 많아짐에 따라 C_G_N과 G_N의 성능은 높아지나 P_Corr, Content를 이용한 방법은 큰 차이를 없음을 나타낸다. 예측의 정확도는 내용 기반 여과와 협력적 여과를 병합한 C_G_N의 방법이 G_N의 방법보다 우세함을 알 수 있다.

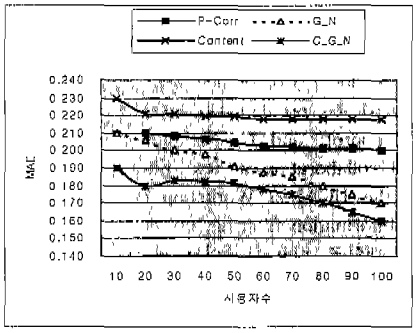

그림 4 사용자 수의 변화에 따른 MAE

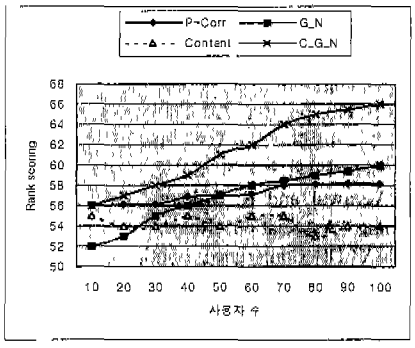

그림 5 사용자 수의 변화에 따른 순위 측정 척도

표 12는 식 (12)와 식 (14)를 기반으로 상품에 대해 평가한 횟수를 증가시킴에 따른 제안된 협력적 여과와 내용 기반 여과를 병합한 추천 방법(C_G_N), Soboroff 방법, Pazzani 방법, Fab 방법의 MAE와 순위 스코어를 나타낸다.

표 12 n번째 평가 횟수에 따른 MAE와 순위 스코어

n번째 평가	MAE				Rank scoring			
	Pazzani	Soboroff	Fab	C_G_N	Pazzani	Soboroff	Fab	C_G_N
10	0.200	0.300	0.230	0.199	57.6	57.4	61.5	56.0
20	0.190	0.280	0.221	0.190	57.6	57.3	61.3	56.5
30	0.190	0.230	0.221	0.186	57.7	58.1	61.5	57.0
40	0.180	0.198	0.219	0.182	57.8	58.5	61.6	57.5
50	0.179	0.191	0.219	0.181	57.8	58.8	61.8	57.8
60	0.178	0.188	0.218	0.178	57.9	59.1	61.9	58.1
70	0.178	0.185	0.216	0.175	58.0	59.8	61.9	58.2
80	0.177	0.181	0.215	0.170	59.0	63.0	62.1	58.6
90	0.176	0.173	0.213	0.165	59.0	68.0	62.1	59.0
100	0.176	0.174	0.215	0.160	60.0	70.0	63.0	59.9

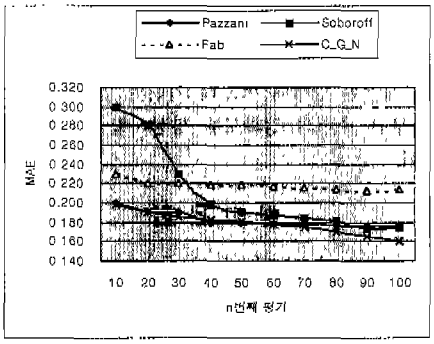

그림 6 n번째 평가에서의 MAE

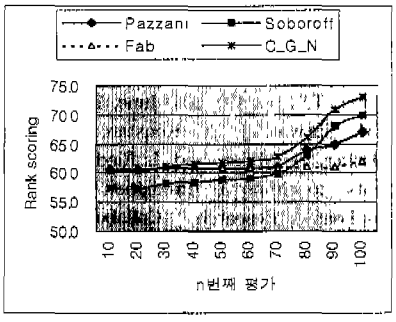

그림 7 n번째 평가에서의 순위 스코어

그림 6과 그림 7은 표 12를 기반으로 한 사용자가 평가한 횟수를 증가시킴에 따른 MAE와 순위 스코어 척도를 나타낸다. 그림 6과 그림 7에서 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 C_G_N의 성능은 전반적으로 높은 예측 정확도를 보이며 그 중 C_G_N은 가장 높은 정확도를 보인다.

6. 결 론

본 논문에서는 협력적 여과 기법에서의 사용자-상품 행렬의 차원수를 감소시키기 위하여 베이지안 분류를 사용하여 상품을 클래스별로 분류하며, 사용자를 대상으로는 유전자 알고리즘을 사용하여 분류된 상품을 기반으로 사용자를 군집시켰다. 또한, 새로운 사용자에 대해 연관 피드백을 사용하여 프로파일을 생성함으로써 초기 평가 문제를 해결하였다. 사용자의 선호도는 차원수가 감소된 협력적 여과를 통한 그룹과 연관 피드백을 통해 생성된 사용자 프로파일의 병합을 통하여 발견된다. 제안된 방법의 성능을 평가하기 위해 기존의 협력적 여과 시스템과 내용 기반 여과를 병합한 방법과 비교한 결과 기존의 방법보다 높은 성능을 보였다.

향후, 제안된 방법에 메모리 기반의 방법을 병합한 추천 시스템은 추천의 정확도를 향상시킬 수 있을 것이라 기대한다.

참 고 문 헌

[1] 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.

[2] R. Raymond and J. Mooney and L. Roy, Content-Based Book Recommending Using Learning for Text Categorization, Proceedings of the Fifth ACM Conference on Digital Libr*aries*, San Antonio, TX, pp. 195-204, June, 2000.

[3] B. M. Sarwar, J. A. 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.

[4] D. M. Pennock and E. Horvitz, Collaborative Filtering by Personality Diagnosis: A Hybrid Memory and Model-Based Approach, Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence, 2000.

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

[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] 인하대학교, 사용자 중심의 지능형 정보 검색 시스템, 최종 연구 개발 보고서, 정보통신부, 1997.

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

[14] 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.

[15] 고수정, 이정현, 연관 단어 마이닝을 이용한 특징추출, 한국정보과학회 논문지 심사중, 2001.

[16] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," Proceedings of the 20th VLDB Conference, Santiago, Chile, 1994.

[17] R. Agrawal and T. Imielinski and A. Swami, "Mining association rules between sets of items in large databases," In Proceedings of the 1993 ACM SIGMOD Conference, Washington DC, USA, 1993.

[18] T. Michael, *Maching Learning*, McGraw-Hill, pp. 154-200, 1997.
[19] 고수정, 이정현, Apriori-Genetic 알고리즘을 이용한 베이지안 자동 문서 분류, 한국정보처리학회 논문지(B), 제8권, 제3호, 2001.
[20] 고수정, 최준혁, 이정현, 연역적 유전자 알고리즘을 이용한 연관 단어 지식베이스의 최적화, 한국정보과학회 논문지(B), 제28권, 제8호, 2001.
[21] 백준호, 최준혁, 이정현, 한국어 웹 정보검색 시스템의 정확도 향상을 위한 연관 피드백 에이전트, 한국 정보처리학회 논문지, 제6권, 제7호, pp. 1832-1840, 1999.
[22] 이수정, 권혜련, 김은주, 이일병, 유전자 알고리즘을 이용한 군집화 기법의 적합도 함수에 관한 연구, 한국정보과학회 '2001춘계학술발표대회 논문집, 2001.
[23] 한승희, 이재윤, 문헌 클러스터링을 위한 유사계수간의 연관성 측정, 제6회 한국정보관리학회 논문집, pp. 25-28, 1999.
[24] M. Gordon, "Probabilistic and genetic algorithms for document retrieval," Communication of the ACM,31, pp. 1208-1218, 1988.
[25] V. Rijsbergen and C. Joost, *Information Retrieval*, Butterworths, London-second edition, 1979.
[26] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, 1983.
[27] 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(이학사). 1997년 인하대학교 교육대학원 전자계산교육(교육학석사). 2000년 인하대학교 대학원 전자계산공학과 박사과정 수료. 1990년 ~ 1991년 수협중앙회 전자계산소 근무. 1992년 ~ 2000년 문성여자정업고등학교 교사. 2001년 ~ 현재 유니버셜 소프트정보통신(주) 주임 연구원. 관심분야는 데이타마이닝, 정보검색, 기계학습

김 진 수

1998년 인천대학교 전자계산학과(공학사). 2001년 인하대학교 전자계산공학과(공학석사). 2001년 ~ 현재 인하대학교 전자계산공학과 박사과정. 관심분야는 자연어처리, 데이터마이닝, 정보검색

김 태 용

1992년 인천대학교 전자계산학과(공학사). 1995 인하대학교 전자계산공학과(공학석사). 1997년 ~ 2000년 인하대학교 전자계산공학과 박사과정 수료. 1995년 ~ 1997년 (주)현대정보기술 정보기술연구소 선임연구원. 1998년 ~ 현재 문경대학 인터넷정보계열 전임강사. 관심분야는 자연어처리, 데이터 마이닝, 정보검색

최 준 혁

1990년 2월 경기대학교 전자계산학과(이학사). 1995년 2월 인하대학교 전자계산공학과(석사). 2000년 8월 인하대학교 전자계산공학과(박사). 1997년 ~2001년 현재 김포대학 컴퓨터계열(소프트웨어개발전공) 조교수. 관심분야는 정보검색, 자연어처리, 지능형 알고리듬

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 전자계산공학과 교수. 관심분야는 자연언어처리, HCI, 정보검색, 음성인식, 음성합성, 계산기 구조