협업 여과 기반의 교육용 컨텐츠 추천 시스템 설계

이용준† · 이세훈†† · 왕창종

요 약

협업여과는 흥미 있어하는 제품이나 개인화된 자료, 항목을 제공하기 의해 전체 집단의 의견을 반영하는 전자상거래에서 일반적으로 이용되는 기술이다. 협업여과는 정확하고 신뢰할 수 있는 도구로 입증되어 여러 분야의 전자상거래 영역에서 활용되고 있으나 아직까지 교육분야에는 한정적으로 적용되고 있다. 본 논문에서는 교육용 컨텐츠 추천에 사용자의 평가 점수를 이용하는 협업여과 방식의 추천시스템을 설계하였으며, 사용자 정보를 이용하여 추천의 정확도를 향상시키기 위한 유사도 보정기법을 도입하였다. 평균절대오차(MAE)와 반응자작용특성(ROC)값을 이용하여 제안한 시스템이 기존의 협업여과방식보다 추천 효율이 우수함을 검증하였다.

The Educational Contents Recommendation System Design based on Collaborative Filtering Method

Yong-Jun Lee† · Se-Hoon Lee†† · Chang-Jong Wang

ABSTRACT

Collaborative Filtering is a popular technology in electronic commerce, which adopt the opinions of entire communities to provide interesting products or personalized resources and items. It has been applied to many kinds of electronic commerce domain since Collaborative Filtering has proven an accurate and reliable tool. But educational application remain limited yet. We design collaborative filtering recommendation system using user's ratings in educational contents recommendation. Also We propose a method of similarity compensation using user's information for improvement of recommendation accuracy. The proposed method is more efficient than the traditional collaborative filtering method by experimental comparisons of mean absolute error(MAE) and reciever operating characteristics(ROC) values.

1. 서 론

인터넷의 활용이 일반화되고, 원격 교육에 대한 요구가 증대함에 따라 교육분야에도 웹 환경으로의 변화가 급속히 이루어지고 있다[12]. 웹 환경을 통하여 교육 내용 매체에 쉽게 접근하고, 과거의 교육내용을 활용하여 여러 다른 교수방법 중 개개인에게 가장 적합한 방법을 학습자에게 제공하기 위한 노력이 지속적으로 진행되고 있다. 인터넷에 접근하는 강사와 학생들은 이러한 추세를 직접적으로 느끼고 있으며, 학습 연구자들은 수업 활동에서 강사와 학생의 인터넷과 웹 자원의 사용에 크게 고무되고 있다[13]. 인터넷을

† 준 회 원: 인하대학교 전자계산공학부 박사과정
†† 정 회 원: 인하공업전문대학교 컴퓨터정보공학부 교수
논문접수: 2003년 2월 28일, 심사완료: 2003년 4월 10일

이용한 교육분야의 적용은 1) 교육의 질을 향상시키고, 높은 품질의 웹 자원 사용, 2) 컨텐츠의 재사용과 공유, 3) 질문기반의 과학 학습과 같은 교육 재구성의 지원으로 구분된다. 그러나 교실에서 이러한 웹 자료의 사용이 항상 생산적이고 학습과 연계되어 사용되고 있지는 못하다. 대부분의 교사가 일이 많고, 인터넷 교육에 필요한 시간이 부족하며, 본질적으로 범위가 한정되어 있지 않고, 지속적으로 변화하며, 여과되지 않은 인터넷에서의 자료를 찾기 위한 기술이 부족하다[6]. 이는 기존의 내용기반 추천을 기반으로 하는 웹 검색엔진에 이러한 교육의 특성이 반영되어 있지 않아 교육의 효율적인 지원이 미비하기 때문이며, 따라서 이에 대한 대안이 필요하다.

웹은 학습자가 자기 주도적으로 학습에 임하도록 하는 환경을 마련해 줄 수 있는 최적의 도구로 인식되고 있다. 그러나 웹으로 이루어지고 있는 교육사이트나 학습 시스템들은 단지 텍스트, 그림, 사운드 등의 자료만을 제공해 주고 학습을 수행케 하는 것이 대부분을 차지하고 있다. 이를 보완키 위한 자기 주도적인 협동학습 모형을 통한 웹기반의 학습시스템에 대한 연구가 꾸준히 진행되고 있다[1, 2]. 그러나 이 경우에도 학습자가 직접적으로 접하는 각각의 자료기반의 평가가 아닌 웹사이트나 결과물에 대한 전체적인 평가가 주를 이루고 있다.

협업 여과는 대상이 되는 항목(사이트, 개별 컨텐츠 등)에 대하여 다른 사용자의 평가를 기반으로 사용자에게 추천을 생성하는 기술이다. 어떤 정보를 이미 보았거나 경험한 사람들의 행동과 의견을 가지고 그 정보를 아직 보지 못한 사람들에게 그 정보의 가치를 예측하여 주는 시스템으로, 다른 사람들의 평가를 의미적으로 수집하고 분석하여 정보를 찾는 시간을 줄일 수 있다. 즉 먼저 내용을 검토한 교사나, 학습자의 의견을 반영하여 아직 내용을 접하지 않은 교사나 학습자에게 정보를 제공하는 방식이다. 협업 여과는 Goldberg에 의해서 정보검색시스템에 적용하는 것을 시작으로 사무 업무그룹과 같은 폐쇄그룹 사용자간의 정보 공유를 위하여 개발된 TAPESTRY, 유즈넷 사용자와 영화를 위한 익명의 협업 여과 기법을 제시한 GroupLens, 음악 추천을 위한 Ringo와 비디오 추천 시스템 등 다양한 종류의 추천시스템에서 사용되고 있다[7].

아직까지 교육분야에 이러한 협업여과를 이용한 추천기법을 활용하는 경우는 많지 않으나 교육분야에 좋은 영향을 줄 수 있다고 밝히고 있다[13]. 그러나 단순하게 협업 여과를 적용함에 따라 기존의 협업여과의 문제점인 희소성(sparsity)에 대한 문제가 향상되지 못하였다. 희소성 문제는 이웃이 평가한 자료가 부족하여 이웃의 의견을 반영하여야 하는 협업여과 추천 결과가 부정확한 경우를 의미한다. 본 논문에서는 협업 여과 추천 계산의 희소성으로 발생되는 추천의 정확도를 보완키 위해 사용자 정보를 이용한 가상 평가값을 활용하여, 예측의 정확도를 높이는 방식을 제안하고 이를 반영한 시스템을 설계하고 구현한다.

2. 관련 연구

이 장에서는 추천시스템을 크게 비 협업여과 추천과 협업여과 추천으로 구분하여 고찰한다.

2.1 비 협업여과 추천

비 협업여과 추천 방식으로는 인구통계기반 추천, 내용기반 추천, 사례기반 추천 등이 있다.

인구통계정보는 특정 항목을 선호하는 사용자의 유형을 구분하는데 사용될 수 있다. 인구통계기반 추천은 사용자의 성별, 나이, 직업 등과 같은 인구 통계 요소에 의한 사용자 유형별 특징을 분석하여 상품을 추천하는 방법이다. 이 기법은 단순한 형태의 정보 여과 활용 방법으로 인구통계 요소의 선정여부에 따라 다양한 형태의 분석과 추천이 가능하다. 특히 이 기법은 사용자의 피드백 정보가 없이도 상품에 대한 추천이 가능하여 시스템의 초기 구축 단계나 처음 방문한 사용자에 대해서도 적용할 수 있어 협업여과의 초기화 문제를 해결하는 방안으로도 사용되고 있다[11].

내용 기반 여과 추천은 사용자의 프로파일과 내용을 표현하는 특징 사이의 비교를 기반으로 항목을 추천한다. 내용기반 추천기법은 개인의

요구나 개인으로부터 입력된 모든 정보와 상품에 포함된 텍스트 정보를 이용하여 여과하는 방법이다. 이 기법은 사용자 프로파일을 통해 과거 구매나 추천 결과를 쉽게 반영할 수 있는 장점이 있으며 추천 속도가 빠르다. 상품에 적용된 경우 항목과 사용자 선호도 사이의 연관관계를 근간으로 사용자 소비를 위한 항목을 추천한다. 개인화 시스템에 적용된 내용기반 정보여과기법은 정보 검색 분야에 근간을 두고 있고, 많은 정보 검색 기술에 채택되었다. 주로 텍스트 문서를 추천하기 위해 사용자의 프로파일과 문서내용을 비교한다. 순수한 내용기반 추천시스템은 사용자가 과거에 등급(rating)을 설정한 항목의 내용을 분석함으로써 생성된 사용자 프로파일을 기반으로 추천이 이루어진다. Infofinder, Newsweeder등은 내용기반 정보 여과 방법에 기반을 둔 개인화 추천 시스템의 예이다. 이러한 내용기반 정보 여과 방법은 몇 가지 문제점을 가지고 있다. 첫 번째는 내용을 분석할 수 있는 정보 종류의 제한성으로 자동 추출이 가능한 텍스트 형태가 대부분이다. 현재의 특징 추출(feature extraction)기술은 몇몇 영역의 항목(예: 영화, 음악, 식당)에는 적용되기에 적합하지 않다. 이러한 항목들은 자동으로 특징을 추출하기가 어렵기 때문이다. 다양한 멀티미디어 형태로 구성되는 교육용 컨텐츠도 이 부류에 속한다. 두 번째는 시스템이 사용자 프로파일에 의해 높은 점수를 얻은 항목을 추천할 경우, 사용자는 그들이 과거에 관심을 보였던 유사한 항목만을 추천 받는다는 것이다[5, 10, 11].

사례기반 추론은 인간이 당면한 문제를 해결하기 위하여 사용하는 추론방법 중 자주 사용하는 방법으로 과거에 이미 해결했던 문제들 중 현재 직면한 문제와 가장 유사한 문제를 기억해 내어 그 문제의 해답에 약간의 수정을 가한 후, 새로운 문제의 해답으로 사용하는 것이다. 사례기반 추론은 이러한 인간의 문제 해결 방식을 모방한 기계 학습(Machine learning) 기법중의 하나이다. 즉, 사례기반추론은 과거에 한번 발생한 문제는 또 다시 비슷한 형태의 문제로 발생할 가능성이 높으며 새로운 문제의 해답 역시 과거의 것과 유사할 것이다라는 가정에서 시작한다. 사례기반추론은 해를 구하고자하는 새로운 사례가 입력되면 저장되어 있는 이전 사례들로부터 비슷한 사례를 검색하고, 적합한 형태의 응답으로 적용시켜 해를 도출한다. 그리고 도출된 해를 다시 저장함으로써 다음의 입력 사례에 대해 더욱 우수한 해를 제시해 줄 수 있도록 재사용 된다. 사례기반추론 기법은 협업 여과와 달리 유사집단의 평가정보를 이용하지 않고, 개인별 속성에 대한 가중치와 속성 값을 이용하여 정보를 추천할 수 있다. 또한 이 기법은 사용자가 선호하는 컨텐츠 유형에 대한 속성을 추천에 반영하는 장점이 있으며, 사용자 프로파일을 이용하여 개인화 된 추천이 가능하나 사례구축이 어렵다는 단점이 있다[4].

2.2 협업여과(Collaborative Filtering) 추천

협업 여과는 다른 사용자의 평가를 기반으로 사용자에게 추천을 생성하는 기술이다. 어떤 정보를 이미 보았거나 경험한 사람들의 행동과 의견을 가지고 그 정보를 아직 보지 못한 사람들에게 그 정보의 가치를 예측하여 주는 시스템으로, 다른 사람들의 평가를 의미적으로 수집하고 분석하여 정보를 찾는 시간을 줄일 수 있다. 이 기법은 숨어 있는 선호 패턴을 발견할 수 있으며, 추천의 정확도가 높은 특징으로 인해 널리 사용될 수 있다. 협업여과 방법은 내용에 대한 분석이 아니라 학습자의 평가를 직접적으로 입력하거나, 사용자의 웹의 접속시간, 클릭횟수, 웹 화면의 저장이나 출력 여부와 행동양식을 점수화하여 학습자간의 유사성에 기준으로 평가 값을 예측하는 경우에도 활용이 가능하다. 협업여과추천은 다음

<표 1> 사용자의 자료에 대한 평가

자료 번호	사용자			
	김	이	박	최
1	1	4	2	2
2	5	2	4	4
3			3	
4	2	5		5
5	4	1		1
6	?	2	5	?

과 같은 과정을 걸쳐 평가를 예측한다. 예를 들

어 6개의 자료에 대한 사용자의 평가가 <표 1>과 같다고 하자. 빈칸은 아직 사용자가 자료를 보지 않은 것을 의미하며, 물음표는 자료를 보았으나 평가를 하지 않은 경우를 의미한다. 평가 결과는 1 ~ 5의 숫자로 표시하며, 5는 매우 좋은 경우, 1은 좋지 않은 경우를 나타낸다.

일반적으로 두 사용자간의 유사도 계산에는 피어슨 상호관계성식과 벡터유사성식이 가장 널리 사용되고 있다[7].

Correlation :

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

Vector similarity :

$$w(a,i)=\sum_j\frac{v_{a,j}}{\sqrt{\sum_{k\in I_a}v_{a,k}^2}}\frac{v_{i,j}}{\sqrt{\sum_{k\in I_i}v_{i,k}^2}} \quad (2)$$

식(1)을 이용하여 김과 이의 유사도를 계산해 보면 다음과 같다.

$$w(\text{김},\text{이})=\frac{\sum_j(v_{\text{김},j}-\overline{v_{\text{김}}})(v_{\text{이},j}-\overline{v_{\text{이}}})}{\sqrt{\sum_j(v_{\text{김},j}-\overline{v_{\text{김}}})^2}\sqrt{\sum_j(v_{\text{이},j}-\overline{v_{\text{이}}})^2}}$$

$$=\frac{-2\ \ -2\ \ -2\ \ -2}{\sqrt{10}\ \sqrt{10}}=-0.8$$

상관계수가 1이면 완전 정상관 관계라 하며, -1 이면 완전 역상관 관계라 한다. 상관관계가 0 이면 상관관계가 없음을 나타낸다. 이와 같은 방식으로 사용자간의 상관관계를 계산해 보면 박과의 관계계수는 1이며, 최와의 상관관계는 0이 된다. 즉 사용자 김과 가장 유사한 이웃은 박, 이, 최의 순이다. 사용자 김이 학습은 하였으나 평가하지 않은 6번째 자료에 대한 평가는 식 (3)을 이용하여 평가를 예측한다.

$$P_{a,i}=\overline{r_a}+\frac{\sum_{u=1}^{n}(r_{u,i}-\overline{r_u})*w_{a,u}}{\sum_{u=1}^{n}w_{a,u}} \quad (3)$$

평가의 예측에는 상관계수를 포함한 평균값을 이용한다.

$$P_{\text{김},6}=\overline{r_{\text{김}}}+\frac{\sum_{u=1}^{n}(r_{u,6}-\overline{r_u})*w_{\text{김},u}}{\sum_{u=1}^{n}w_{\text{김},u}}$$

$$=3+\frac{2r_{\text{김박}}-r_{\text{김이}}}{|r_{\text{김박}}|+|r_{\text{김이}}|}=3+\frac{2-(-0.8)}{|1|+|-0.8|}=4.56$$

위의 예에서 n은 평가자의 수를 의미한다. 최와의 상관관계는 0 임으로 식 (3)의 계산에서 최와의 관계 값은 제외되어 김의 6번째 자료에 대한 평가 값은 4.56으로 예측된다. 위와 같은 방법으로 평가하지 않은 자료에 대한 예측을 수행한다. 위의 방법은 협업여과의 일반적인 과정을 적용한 예를 보인 것이다. 협업여과는 여러 분야에서 활용되고 있으며 좋은 결과를 보이고 있다. 그러나 1) 자료에 대한 초기 평가가 없는 경우에는 추천이 불가하며, 2) 항목의 전체 수에 비해서 사용자들이 평가한 자료의수가 적은 경우 추천의 정확성이 떨어지는 문제점이 지적되고 있다[7].

이러한 협업여과의 단점을 개선하기 위한 다양한 방법들이 연구되었다.

Pazzani[11]는 사용자의 프로파일을 가중치가 부여된 단어의 집합으로 표현하였다. 예측은 직접 내용-프로파일의 행렬에 협력적 여과를 적용함으로써 이루어진다. 여기서 내용-프로파일 행렬은 여러 사용자들의 프로파일의 모임이다.

Balabanovic[5]의 Fab는 사용자의 연관 피드백과 "주제" 여과를 통한 내용 기반 여과를 하며, 이를 협력적 여과와 병합하는 방법을 제안하였다. 내용 기반 여과에서 문서는 주제 여과에 의해 여과되어 문서에 대한 순위 목록을 만들며, 생성된 목록에 대해 사용자는 연관 피드백을 제공함으로써 여과가 이루어진다.

Melville[10]는 빈약한 사용자의 평가 행렬을 내용기반예측을 통해 모의(pseudo) 사용자 평가 행렬을 생성하고, 이를 기반으로 협업여과 추천을 진행한다. 정확도가 크게 향상되지는 못하였다.

2.3 교육분야의 협업여과 추천

교육용 컨텐츠를 체계적으로 분류, 저장, 검색

하기 위한 방안으로 메타데이터를 확장하여 이용한 방식이 디지털 도서관을 중심으로 활발하게 진행되고 있다[9]. 또한 규격화 된 메타데이터를 확장과 함께, 비 규격화된 사용자의 선호도나 흥미 등을 반영하는 방안도 연구되고 있다. Recker[12]는 협업여과방법을 교육용 컨텐츠 추천에 활용하여, Altered Vista (www.alteredvista.usu.edu)라고 불리는 협업여과방법을 이용한 교육 시스템을 설계하고, 개발하여 학습환경과 참여자를 포함한 실험적 연구를 진행하여 협업여과가 멀티미디어를 기반으로 하는 교육용 컨텐츠 활용에 효과가 있음을 실험하였다. 교육내용에 대하여 몇 가지 항목(유용성, 교육 관련성, 품질, ..., 총 평가점수)을 1 ~ 5의 점수로 부여하여 사용자간의 유사도를 계산하는 방식을 도입하였다. 또한 개인별 추천 시스템을 도입치 않고 일반적인 평균값을 기반으로 추천하는 방식과 개인별 추천 시스템을 도입한 경우를 비교하여 개인별 추천 시스템을 도입한 경우가 보다 효율적이었음을 검증하였다. 그러나 이 경우에도 사용자들의 점수 부여 기피로 협업여과의 희소성 문제를 향후 해결하여야 할 문제점으로 지적하였다. 또한 초기 시스템에서 시도한 여러 항목에 대한 평가점수 부여 방식이 희소성이 높아 효율성이 낮음에 따라, 단일 항목에 대한 평가점수 방식으로 시스템을 변경하였다[13].

3. 교육용 컨텐츠 추천 시스템

3.1. 시스템 설계

웹 자원을 사용하는 교수자와 학습자가 대상이 되어 웹 자원을 평가하고 점수를 부여하여, 다른 사용자의 추천에 도움을 줄 수 있다. 사용자와 선호도가 유사한 이웃을 찾을 수 있다면 자기 주도적 개별학습의 효율을 높일 수 있을 것이다. 이러한 자료가 누적되면 시스템으로부터 개인화된 추천도 요청할 수 있다. 별 필요 없는 자료는 거부하는 대신, 가치 있는 정보를 지정하여 다른 사람의 의견을 가름해 볼 수 있다. 이와 같이 사용자 모델을 기반으로 하는 협업여과를 이용한 시스템을 구축하려면 먼저 설계 시 몇 가지 고려하여야 할 사항이 있다.

첫 번째는 점수 부여를 받기 위한 화면의 설계이다. 점수 부여 항목은 단순하게 작성하여 평가자가 쉽게 평가가 진행될 수 있어야 한다. 일반적으로 사용자들은 평가를 하지 않고 무임승차하려는 경향이 있다. 따라서 평가에 대한 대가를 보장할 수 있다면 보다 많은 사용자의 협조를 얻을 수 있다.

두 번째는 평가 점수의 취득에 관한 부분이다. 사용자를 통하여 직접 점수를 부여받는 방법은 정확하고 빠르나, 사용자가 기피하는 경향이 있어 협업여과의 문제점인 희소성의 원인이 된다. 간접적으로 사용자의 학습형태를 관찰하여, 해당 콘텐츠의 사용 빈도나 머무는 시간, 저장여부, 출력 여부 등을 측정하여 평가에 활용할 수 있다. 그러나 측정에 소요되는 시간이 오래 걸리고 정확성이 떨어지는 단점이 있다.

세 번째는 평가 점수의 집계 방법이다. 일반적으로 모든 평가 자료는 데이터베이스에 저장되어 예측의 기본자료로 활용되며, 건별로 처리하거나, 교육적 특성 등을 고려하여 일괄 입력하는 방안도 고려할 수 있다.

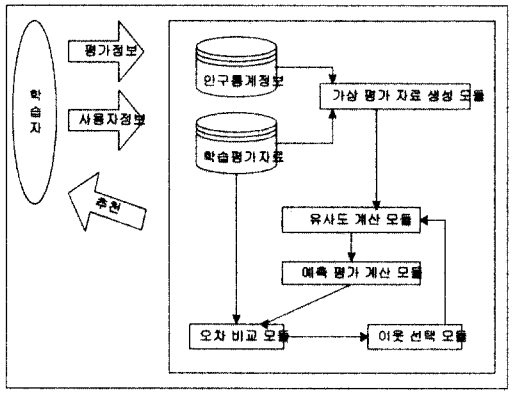
네 번째는 사용자 평가 정보의 활용 방법이다. 사용자가 흥미 있어하는 자료를 내용-색인으로 대체하여 제공할 수 있으며, 모든 평가를 색인어에 의해서 검색할 수 있다. 또한 사용자가 추천을 요구하는 경우 가장 유사한 이웃을 선택하여 제공한다. 적어도 두 개 항목의 평가가 있어야 사용자간의 유사도 계산이 가능하며, 적어도 유사도가 한계치 이상인 경우에 한해서만 추천에 포함시킨다. 또는 최적의 이웃의 수를 지정하여 추천을 진행할 수도 있다. 이 경우 한계치와 이웃의 수는 예측의 신뢰성을 보장하여 주는 중요한 변수로 정확한 예측을 위해서는 정확한 유사도 계산이 필수적이다. 점수가 1 ~ 5 사이로 정의된 경우 예측 결과가 4 이상인 경우만 추천하기도 한다[7].

다섯 번째는 사용자의 확인 문제이다. 사용자는 시스템에 접근하기 위하여 로긴을 하고, 로긴되어 평가된 자료는 평가자가 누구인가를 확인할 수 있다. 일부 시스템에서는 개인 정보 보호 측

면에서 익명을 이용하기도 한다. 로긴한 사용자는 본인이 원하는 영역의 특정 페이지를 검토하고 평가점수를 부여한다. 평가점수는 데이터베이스에 저장되며, 협업여과나 사용자 요구 정보 검색에 사용된다. 또한 예측 정확도 평가를 위해 평가점수는 학습자료와 실험자료로 활용된다.

3.2. 시스템 구성도(System Architecture)

시스템은 (그림 1)과 같으며, 유사도 가상 평가 점수 생성 모듈, 유사도 계산 모듈, 이웃 선택 모듈, 예측 평가 계산 모듈, 오차 비교 모듈로 나뉜다. 가상 평가 점수 생성 모듈은 유사도 보정 기법을 위해 필요한 사용자 모델별(성별, 학력, ..., 학습 능력) 평균값을 산출하고, 학습 평가 자료와 사용자 모델별 평균값을 이용하여 가상 평가 점수를 생성하며, 가상 평가 점수는 유사도 계산에 사용된다. 유사도 계산 모듈은 사용자간의 유사도를 계산한다. 계산시간이 오래 걸리므로 오프라인에서 작업하며, 주기적으로 갱신한다.



(그림 1) 시스템 구성도

이웃선택 모듈은 사용자에 대한 평가 추천에 대한 오차가 커지지 않도록 주기적으로 검사하여 오차가 상대적으로 커지는 경우 오프라인에서 시뮬레이션을 통하여 근접 이웃의 수를 조정한다. 예측 평가 계산 모듈은 유사도를 기준으로 사용자에 대한 추천을 위한 예측 평가 값을 생성한다. 오차 비교 모듈은 계산된 평가 점수와 사용자의 실제 평가 점수를 비교하며 오차가 커지는 경우 이웃 선택 모듈을 통해 근접 이웃의 수를 조정할 수 있는 기본 자료를 제공한다.

알고리즘1 사용자 정보기반 추천 정확도 향상

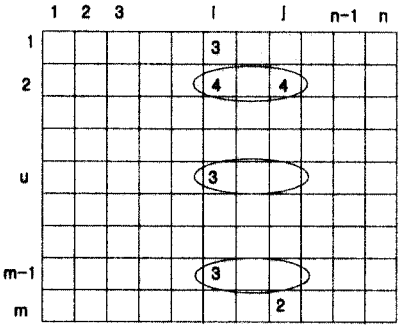
```
//Create environment (off-line-통계정보 생성등)
INPUT : 학습 자료(Train Data)
BEGIN:
 For maxuser
  For maxmovie
   Create Train-Matrix, 각영화별 각 군집별 평균값 계산
   If 사용자 점수 ≠ null and 이웃 점수 = null
     then
        Add demographic 가상 평가값 to train-matrix
   End if;
   Calculate average
  End for;
 End for;
 For maxuser
  For maxmovie
   Calculate correlation
  End for
 End for;
END:
OUTPUT : Correlation-matrix
//Calculation estimate(on-line)
INPUT : 실험 자료(Test Data)
BEGIN:
While (유사도 >= 지정치 or 근접 이웃 >= 설정치)
  Calculate estimation
End while;
  Calculate mean absolute error
END:
OUTPUT : Mean absolute error
```

알고리즘 1에서 보인 바와 같이 인구 통계 정보를 문제 영역에 반영하는 작업(각 항목별, 각 군집의 평균값 계산)을 통하여 문제 영역의 희소성을 감소시킬 수 있도록 하였다.

3.3. 유사도 보정 기법

유사도 보정 기법은 사용자 모델 정보를 정의하여, 일종의 범주화(Clustering)를 통하여, 이웃군의 평균을 형성하고, 유사도 계산 시 필요한 이웃의 빈 평가자리를 채우기 위해 가상 평가 값에 이 이웃군의 평균값을 이용하는 방식이다. (그림 2)의 행은 항목의 종류(m)를, 열은 사용자(n)을 나타내며, 이는 m x n 행렬을 구성하게 된다. 유사도 계산은 상대성이 있어서, 상대되는 항목이 없으면 계산에서 제외가 된다. i 번째 사용자의 m 번째 항목의 추천을 하는 경우 j 번째 이웃과의 유사도 계산에서 u 번째, m-1 번째 항목은 j 번째 사용자의 정보가 없어 제외되게 된다. 이 경우 2번째 항목의 평가만으로 유사도를 계산하

게 되어 정확하게 유사한 이웃인지가 확인되지 않는다. 따라서 정확한 계산을 위해서는 사용자와 이웃 간의 상호성을 보장하여 주어야 한다. 사용자의 자료는 있으나 이웃의 자료가 없는 경우 가상 평가 값으로 보완하고 계산을 수행하면 계산의 정확도가 높아질 것이다. 빈자리를 채우는 가상 평가 값을 어떤 항목을 기반으로 하느냐에 따라 계산 결과는 달라질 수 있다. 유사도 계산 시 사용자 모델을 이용하여, 빈자리에 가상 평가 값을 추가하고, 유사도 계산의 정확도를 향상시키도록 하였으며, 이는 예측 계산의 정확도를 향상시키는 기반이 된다. (그림 2)의 j번째 이웃의 사용자 모델을 기반으로 u 번째, m-1 번째 항목의 j 번째 이웃의 빈자리를 채우고 유사도를 계산하여, 보다 정확도 높은 유사도 계산을 유도하였다.



(그림 2) 사용자-항목 행렬 구성

일반적으로 다음과 같이 이웃군을 이용하여 가상 평가 값을 구성한다.

사용자 모델이 k개의 속성을 가진다면, $P=\{P1, P2, \cdots, Pk\}$ 이다.

예를 들어 k = 3 인 경우 $P=\{P1, P2, P3\}$이다. 여기에서 속성 Pj가 Sj개의 다른 속성을 가지면,

$Pj=\{q_{1,}q_{2,}\cdots, q_{s_j}\}$로 표현된다.

속성 Pj를 n개씩 범주화할 경우 다음과 같은 형태를 갖는다.

$$Pj_1=\{q_{1,}q_{2,}\cdots, q_n\}$$

$$Pj_2=\{q_{n+1}, q_{n+2}, q_{2n}\}$$

$$\cdots$$

$$Pj_m=\{q_{m+1}, q_{m+2}, q_{mn}\}$$

$$\cdots$$

이 범주의 대표 값(평균)을 첫 번째 항목을 선택한 경우

$V_{P1_1}=\dfrac{q_1+q_2+\cdots+q_n}{n}$ 로 나타난다.

예를 들어 이웃 u가 P1의 첫 번째 군에 속하고, P2의 두 번째 속성에 속하며, P3의 첫 번째 속성에 속한다면 이웃 u의 속성인 C_u는 $\{V_{P1_1}, V_{P2_2}, V_{P3_1}\}$로 구성된다. 여러 개의 속성 중 각 개인의 특성에 따라 평가에 반영되는 정도의 차이를 고려하여, 빈자리에 적용하는 가상 평가 값을 다음과 같이 정의하였다.

$$S_{u,i}=E_1*V_{P1_1}+E_2*V_{P2_2}+E_3*V_{P3_1} \qquad (4)$$

$$\text{where } \sum_{j=1}^{k}E_j=1,\quad 0\le\forall E_j\le 1$$

u 는 대상이웃, i는 대상 항목, E_j는 가중치

E_j는 오프라인에서 시뮬레이션을 통하여 최적치를 계산하며, 주기적으로 갱신하여 반영토록 하였다. 이 경우에도 학습 자료가 적으면 속성별 군집 대표 값 V_{Pj}에서 희소성이 발생하게 된다. 이러한 희소성은 군집 테이블내의 근접 자료를 활용하여 희소성을 감소시켰다.

$$V_{Pj}=\frac{V_{P_{j-1}}+V_{P_{j+1}}}{2} \quad \text{if } V_{Pj}=\phi \qquad (5)$$

유사도 계산을 위한 피어슨 상관관계식 (6)에서 $r_{a,i}$은 있으나, $r_{u,i}$ 가 없을 경우 $r_{u,i}$는 가상 평가 값인 $S_{u,i}$값으로 대체된다.

$$w_{a,u}=\frac{\sum_{i=1}^{m}(r_{a,i}-\overline{r_a})*(r_{u,i}-\overline{r_u})}{\sigma_a*\sigma_u} \qquad (6)$$

$w_{a,u}$는 사용자 a와 이웃 u간의 유사도 가중치이다. 계산된 유사도를 기반으로 지정된 근접 이웃의 수를 참조하여 사용자가 원하는 대상의 평가 예측치를 식 (7)을 이용하여 계산한다.

$$P_{a,i}=\overline{r_a}+k_{weight}\frac{\sum_{u=1}^{n}(r_{u,i}-\overline{r_u})*w_{a,u}}{\sum_{u=1}^{n}w_{a,u}} \qquad (7)$$

많은 평가를 한 사용자가 적은 평가를 한 사용자 보다 신뢰도가 높다는 점을 고려하여, 평가를 적게한 사용자와 많이 한 사용자를 차별하기 위하여 식(7)에 중요도 가중치 K-weight를 반영

하였다[7].

$$k_{weight}=\frac{I_i}{m}: \text{if } I_i<m,\ else\ I_i=1 \quad (8)$$

I_i : 전체 항목 중 평가된 항목의 수

m : 이웃의 수

최종적으로 계산된 예측치와 실측치와의 차이를 계산하여 평가의 정확도를 비교한다.

4. 실험 및 평가

실험은 16개 시도 교육청 공유 체재[3]에 등재되어 있는 140건의 교육자료(파워포인트 형태)를 이용하여 실험을 하였다. 40명의 초등학교 5학년 학생을 대상으로 학기말에 그동안 공부하였던 항목의 자료를 평가하게 하였다. 실험 환경은 Window 2000 서버와 MS-SQL 데이터베이스를 이용하였고, 평가자가 평가 결과를 입력한 결과의 유사도 계산등은 COM 컴포넌트로 구현하였다. 취득된 데이터 집합은 1 ~ 5 사이의 점수로 평가된 3,214개의 학습 자료 평점으로 최소 30개 이상의 학습자료를 평가한 29명의 사용자가 본 3,093개 학습 평가 자료를 대상으로 구성되어 있다. 사용자의 정보로는 성별, 과학성적, 과학관심도, 전체성적, PC 능숙도, 교우관계, 부친직업 등이 포함 되어있다. 총 3,093개의 데이터 집합 중에서 학습자료로 2,400개의 정보가 실험자료로 693개의 정보로 구분되어 있다. 사용자 정보는 다음과 같이 구성된다.

P = {성별, 과학성적, 과학관심도,..., 부친직업}

P11 = {남}, P12 = {여}

P21 = {3}, P22 = {2}

...

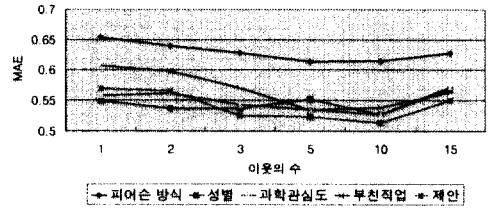
P71 = {회사원}, P72 = {대학교수, 의사, 변호사}, P73 = {가공업, 요식업,..., 상업}

(그림 2)에서 비어 있는 j번째 이웃의 u번째 항목을 가상 평가값으로 대체하는 방법은 다음과 같다. j번째 이웃의 인구 통계 정보가 남자이고, 과학성적이 3 이며, 과학관심도가 3이고, ... 부친직업이 회사원이라면, u번째 자료에 대한 남자들이 평가한 점수의 평균값과 과학성적이 3인 학습자가 평가한 u번째 자료에 대한 평균값, ... 부친의 직업인 학습자가 평가한 u번째 자료에 대한 평균값을 계산하고, 이 결과를 가상 평가값으로 사용한다. 각 자료의 각 군집에 대한 평균값은 Train-Matrix 구성 시 오프라인 작업으로 미리 구성한다.

본 논문에서는 실제 부여 점수와 예측간의 차이 분석에 많이 사용되고 있는 평균에러(mean absolute error)방법[10]을 사용하여 제안한 방법의 타당성을 검증하였다.

$$\text{평균에러}(MAE)=\frac{|R_1-P_1|+\cdots+|R_n-P_n|}{n} \quad (6)$$

(그림 3)은 사용자 정보인 성별, 과학관심도, 부친직업을 각각 반영한 경우와 3항목을 함께 반영한 제안 시스템의 경우를 나타낸 것이다. 어느 경우에도 결과치가 피어슨 상관관계를 이용한 방식보다 좋음을 알 수 있다. 즉 희소성을 감소시켜 주면 효과가 있음을 알 수 있다.



(그림 3) 사용자 정보를 반영한 실험결과

이웃의 수에 따라 예측 결과의 정확도가 크게 차이가 나며, 대상이 되는 영역이나, 학습 평가 자료의 크기에 따라 크게 차이가 나므로, 학습 평가 자료를 기준으로 최적의 이웃의 수를 선정하고 이 이웃의 수를 예측 평가시의 이웃의 수로 이용한다.

<표 2> 사용자 정보를 반영한 실험 결과 (이웃의 수 n = 10인 경우)

구분	MAE	ROC-3			
		Sensitivity	Specificity	Accuracy	Error rate
피어슨 방식	0.6149	0.7061	0.2535	0.5477	0.4522
성별반영	0.5268	0.9072	0.4401	0.7437	0.2562
과학관심도 반영	0.5254	0.8943	0.4258	0.7303	0.2696
부친직업반영	0.5385	0.9072	0.4449	0.7451	0.2548
제안	0.5184	0.9201	0.4210	0.7453	0.2546

<표 2>는 ROC(Receiver Operating Character)

측정에 대한 비교이다[7]. 전체 자료의 평균값이 3 주변이어서 ROC-3으로 비교하였다. ROC-3에서는 사용자의 평가 정보 중 3, 4, 5의 값은 좋은 값(positive)으로, 1, 2의 값은 나쁜값(negative)으로 정의한다. Sensitivity는 임의로 선택된 평가값이 좋은 값으로 추천될 확률로, 그 값이 1인 경우 완벽한 경우이며, 0.5인 경우 무작위(random)로 판별한다.[7] Specificity는 임의로 선택된 평가값이 나쁜값이 추천되지 않을 확률이다. Accuracy는 전체 실험 자료 중 예측이 맞은 경우의 확률이며, Error rate는 전체 실험 자료 중 예측이 틀린 경우의 확률이다. 사용자 정보를 반영한 경우 기대 했던 바와 같이 정확도가 좋아짐을 확인할 수 있었다. 정보를 적용할 것인가를 선택하는 작업은 문제 영역에 종속적이어서 매우 어렵다. 본 실험에서는 7가지의 사용자 정보 중 실험 반영 시 영향력이 큰 3개 항목만을 제안 시스템에 반영하였다. 실험에서는 오프라인에서 계산을 통하여 오차가 작은 항목을 제안방식에 적용하였으나, 사용자 정보의 선택이나 사용자 정보 중 어떤 항목을 사용할 것인가는 보다 심도 있는 연구가 진행되어야 할 부분이다.

5. 결 론

본 논문에서는 기존의 협업여과 방법을 개선키 위하여 유사도 보정 기법을 포함하는 교육용 컨텐츠 추천 시스템을 설계하여 기존의 협업여과 방식보다 우수함을 실험을 통하여 검증하였다. 상관관계를 이용하는 유사도 계산식은 비교 대상의 상호성에 대한 자료가 확보되어 있는 경우 보다 정확한 결과를 얻을 수 있다. 따라서 상호성이 보장될 수 있도록 가상 평균값을 활용하여 유사도 계산 정확도를 높여, 결과적으로 예측 결과를 높일 수 있음을 확인하였다. 또한 첫번의 추천을 이용하는 사용자의 평점 결과가 없는 상태에서도 사용자 정보를 활용하여 추천이 가능하므로 처음 사용자에 대한 초기화에 대한 문제도 해결이 가능하다. 실험에서 보인 바와 같이 사용자 정보를 보완하면 보다 높은 평가예측의 가능성이 있음을 확인하였다. 그러나 유사도 보완을 위하여 어떤 성격의 자료를 사용하는 것이 효율적인지는 보다 많은 실험과 연구가 지속적으로 진행되어야 할 부분이다.

참 고 문 헌

[1] 김효준 · 조세홍(2001). 자기 주도적인 협동학습 모델을 통한 웹(Web) 기반 학습시스템 설계 및 구현. 컴퓨터교육학회 논문지 5(1).

[2] 서원석 · 김현철 · 이원규(2001). 학습자간의 상호작용 강화를 위한 웹 기반 협동학습의 구현 및 적용. 컴퓨터교육학회 논문지 5(4).

[3] 최종연구개발보고서(2001). 범국가적 교육단위기관간 교육용 컨텐츠 공동활용방안연구. 경상남도교육청, 인천광역시교육청, (주)액티브웹 기술연구소, 인하대학교부설 컴퓨터과학응용연구소.

[4] Aamodt,A. and Plaza,E(1994), Case-based Reasoning:Foundational Issues, *Methodological Variations, and System Approaches, Artificial Intelligence Communications*, Vol.7, No.1. 39-59

[5] Balabanovic, M., and shoham, Y.(1997), Fab : Content-based, collaborative recommendation, *Communications of the Association of Computer Machinery* 40(30), 66-72.

[6] Borgman,D.,Krieger,D., Gallagher, Al,&Bower,J(1990). Children's use of an interactive science library :Exploratiory research, School Library Media Quaterly, 18, 108-113

[7] Herlocker,J., Konstan,J., Borchers, A.,Riedl,J(1999), An Algorithmic Framework for Performing Collaboraive Filtering, *Proc. of the Conference on Research and Development in Information Retrieval*. 203-237.

[8] Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, K., and Riedl, J(1997), GroupLens:Applying Collaborative Filtering to Usenet News, *Communications of the ACM*, 40(3), 77-87.

[9] LTSC(2000), IEEE P1484.12 Learning Objects Metadata Working Group(http://

ltsc.ieee.org.wg12/index.html)

[10] Melville,P., Mooney,R., Nagarajan, R(2002), Content-Boosted Collaborative Filtering for Improved Recommendations, *Proc. of the eighteenth National Conference on Artificial Intelligence*, 187-192.

[11] Pazzani,M.J.(1999) ,A Framework for Collaborative, Content-Based and Demographic Filtering, Artificial Intelligent Review, 394-408.

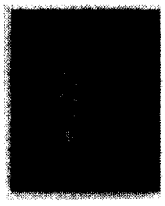
[12] Recker,M., Wiley, D.(2000), An interface for collaborative filtering of educational resources, *Proc of the International Conference on Artificial Intelligence*, 317-323.

[13] Recker,M., Walker, A.(2002), What do you recommend ? Implementation and analysis of collaborative filtering of Web resources for education, *Proc. of International Conference on Artificial Intelligent*, http::/it.usu.edu/~mini/papers/instscience1.doc 2002.

이 용 준

1982 인하대학교 전자계산학과(이학사)
1995 인하대학교 전자계산학과(공학석사)
2001 인하대학교 컴퓨터공학부 박사 수료
1982~2000 한국전기연구원 선임연구원
1996~1999 한국전기연구원 전산실장
1999~현재 한국전기연구원 전기시험연구소 선임기술원
관심분야: e-Learning, 소프트웨어공학, 분산객체 컴퓨팅, DAS
E-Mail: yjlee@keri.re.kr

이 세 훈



1985 인하대학교 전자계산학과(이학사)
1987 인하대학교 전자계산학과(이학석사)
1996 인하대학교 전자 계산공학과(공학박사)
1987~1990 해병대 분석 장교
1990~1993 (주)비트컴퓨터 기술연구소 선임연구원
1999.5 멀티미디어기술사
2001~2002 미국 뉴저지 공과대학(NJIT) 교환교수
1993~현재 인하공업전문대학 컴퓨터정보공학부 교수
관심분야: e-Learning, 하이퍼미디어시스템, 소프트웨어공학, 분산객체컴퓨팅, XML/JAVA
E-Mail: seihoon@inhatc.ac.kr

왕 창 종

1964년 고려대학교 물리학과(이학사)
1975년 성균관대학교 대학원
1985~1992 한국정보과학회 이사회 임원
1993 한국정보과학회 부회장
1979~2003. 2 인하대학교 공과대학 컴퓨터공학부 교수
관심분야 : 소프트웨어공학, 분산객체기술, 컴퓨터기반교육.
E-Mail: cjwang@inha.ac.kr