협업 여과의 희소성을 개선한 교육용 컨텐츠 추천 시스템

이용준° 이세훈 왕창종

°인하대학교 컴퓨터공학부 인하공업전문대학 컴퓨터정보공학부 인하대학교 컴퓨터공학부

yjlee@keri.re.kr, seihoon@inhatc.ac.kr, cjwang@inha.ac.kr

Improving Sparsity Problem of Collaborative Filtering in Educational Contents Recommendation System

Yongjun Lee° Sehoon Lee ChangJong Wang

School of Computer Science & Engineering, Inha University

School of Computer & Information Systems, Inha Technical College

요 약

본 논문에서는 교육용 컨텐츠 추천시스템의 정확도를 향상시키고자 사용자 모델 정보를 활용하여 기존의 협업여과 방법의 유사도 계산을 보완함으로써 추천의 정확도를 향상시키는 방법을 제안하고자 한다. 협업여과방법은 사용자의 평가와 비슷한 선호도를 가지고 다른 사용자의 평가를 기반으로 제품이나 항목을 예측하고 이를 사용자에게 추천한다. 그러나 협업여과방법은 일정 수 이상의 상품이나 항목에 대한 평가가 이루어져야 하며, 사용자의 평가가 적은 경우 희소성으로 인한 평가의 정확도가 낮아지는 단점을 기지고 있다. 본 논문에서는 인구 통계 정보를 이용한 가상 평가 점수를 반영하여 유사도 계산시 희소성을 낮춰 예측의 정확도를 향상시키고자 한다.

1. 서 론

인터넷의 활용이 일반화 되고, 원격 교육에 대한 요구가 증대함에 따라 교육분야에도 웹 환경으로의 변화가 급속히 이루어지고 있다. 인터넷에 접근하는 강사와 학생들은 이러한 추세를 직접적으로 느끼고 있으며, 학습 연구자들은 수업 활동에서 강사와 학생의 인터넷과 웹 자원의 사용에 크게 고무되고 있다[1]. 그러나 교실에서 이러한 웹 자료의 사용이 항상 생산적이고 학습과 연계되어 사용되고 있지는 못하다[2]. 대부분의 교사가 일이 많고, 인터넷 교육에 필요한 시간이 부족하며, 인터넷에서 본질적으로 범위가 한정되어 있지 않고, 지속적으로 변화하며, 여과되지 않은 자료를 찾기 위한 기술이 부족하다[3]. 기존의 웹 검색엔진들은 이러한 교육의 특성이 반영되어 있지 않아 교육의 효율적인 지원이 되지 않으므로 이에 대한 대안이 필요하다.

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. 협업 여과는 Goldberg에 의해서 정보검색시스템에 적용하는 것을 시작으로 사무 업무그룹과 같은 폐쇄그룹 사용자간의 정보 공유를 위하여 개발된 TAPESTRY[4], 유즈넷 사용자와 영화를 위한 익명의 협업 여과 기법을 제시한 GroupLens[5], 음악 추천을 위한 Ringo[6]와 비디오 추천[7] 시스템 등 다양한 종류의 추천시스템에서 사용되고 있다. 아직까지 교육분야에 이러한 협업여과를 이용한 추천기법을 활용하는 경우는 많지 않으나 교육분야에 좋은 영향을 줄 수 있다고 밝히도 있다. 그러나 단순하게 협업 여과를 적용함에 따라 기존의 협업여과의 문제점인 희소성에 대한 문제가 향상되지 못하였다[8]. 본 논문에서는 협업 여과 추천 계산의 희소성(sparsity)으로 발생되는 추천의 정확도를 보완하기 위하여 사용자 모델 정보를 이용한 가상 평가 값을 활용하여, 예측의 정확도를 높이는 방식을 제안하고자 한다.

2. 관련 연구

협업여과의 단점인 희소성을 개선하기 위한 다양한 방법들이 연구되었다. Soboroff[9]는 특징-문서 행렬을 생성하고, 이 행렬의 값에 문서에 나타난 특징 빈도에 비례하여 주어진 가중치를 부여한다. 내용 프로파일 행렬을 생성하기 위하여 정의된 특징-문서 행렬과 협력적 여과에 의한 선호도 평가 행렬을 곱하고, 이 행렬의 SVD(Singular Value Decomposition)를 계산한다. LSI(Latent Semantic Indexing)를 사용하여 내용 프로파일 행렬의 순위를 계산한다. 사용자를 표현하는 중앙값은 사용자에게 적합한 문서의 특징 벡터로부터 추출된다. 새로운 문서는 LSI 공간에서 각 사용자의 프로파일을 기준으로 순위가 부여된다. 계산속도 향상의 부가적인 효과를 거둘 수는 있었으나 결과적으로 정확도가 크게 향상되지 못하였다[10].

Basu[7]은 추천을 분류작업으로 간주하였다. Ripper 라는 연역추론 시스템을 사용하여 사용자와 상품과의 관계를 학습하고, 사용자가 상품을 좋아할지 여부를 예측한다. 이진으로만 분류하므로 다양한 형태의 추천에 적용할 수 없다는 단점을 갖는다.

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[12]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.

Good[13]은 여러 개인화 정보 여과 에이전트와 협력적 여과를 병합함으로써 항목에 대한 추천을 제공한다. 새로운 사용자에 대한 예측은 새로운 사용자의 개인화 에이전트를 협력적 여과에 응용함으로써 생성된다. 에이전트를 이용하는 경우는 희소성 보다는 초기화 문제를 해결키 위한 방편으로 이용된 경우가 많다[14].

Melville[15]는 빈약한 사용자의 평가 행렬을 내용기반예측을 통해 모의(pseudo) 사용자 평가 행렬을 생성하고, 이를 기반으로 협업여과 추천을 진행한다. 정확도가 크게 향상되지는 못하였다.

Recker[8]는 협업여과방법을 교육용 컨텐츠 추천에 활용하여, Altered Vista (alteredvista.usu.edu)라고 불리우는 협업여과방법을 이용한 교육 시스템을 설계하고, 개발하여 학습환경과 참여자를 포함한 실험적 연구를 진행하였다. 협업여과가 멀티미디어를 기반으로 하는 교육용 컨텐츠 활용에 효과가 있음을 실험하였다.

3. 유사도 보정 기법

Recker[8]는 교육용 컨텐츠의 추천에서도 희소성 문제가 가장 큰 것으로 언급하고 있다. 유사도 보정 기법은 사용자 모델 정보를 이용하여, 일종의 범주화(Clustering)를 진행하고, 범주화된 이웃군의 평균을 형성하며, 유사도 계산 시 필요한 이웃의 빈 평가자리를 채우기 위해 가상 평가 값으로 이 이웃군의 평균 값을 이용하는 방식이다. 아래 그림의 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 계산하게 되어 정확하게 유사한 이웃인지가

확인되지 않는다. 따라서 사용자의 자료는 있으나 이웃의 자료가 없는 경우 가상 평가 값으로 보완하고 계산을 수행하면 계산의 정확도가 높아질 것이다.

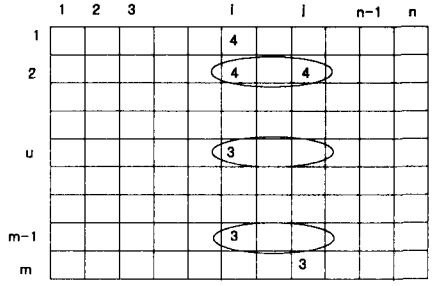

그림 1 사용자-항목 행렬 구성

유사도 계산 시 사용자 모델 정보를 이용하여, 빈자리에 가상 평가 값을 추가하고, 유사도 계산의 정확도를 향상시키도록 하였으며, 이는 예측 계산의 정확도를 향상시키는 기반이 된다. 그림 1의 j번째 이웃의 사용자 모델 정보를 기반으로 u 번째, m-1 번째 항목의 j 번째 이웃의 빈자리를 채우고 유사도를 계산하여, 보다 정확도 높은 유사도 계산을 유도하였다. 일반적으로 다음과 같이 이웃군을 이용하여 가상 평가 값을 구성한다. 사용자 모델 정보가 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 범주화 할 경우 다음과 같은 형태를 갖는다.

$$Pj_1 = \{q_1, q_2, \cdots, q_n\}$$

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

$$\cdots$$

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

이 범주의 대표 값(평균)을 첫 번째 항목을 선택한 경우

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

예를들어 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, 빈 자리에 적용하는 가상 평가 값을 다음과 같이 정의하였다.

$$S_{u,i} = E_1 * V_{P1_1} + E_2 * V_{P2_2} + E_3 * V_{P3_1}$$

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

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.

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

유사도 계산을 위한 피어슨 상관관계식 (3)에서 $r_{a,i}$는 있으나, $r_{u,i}$ 가 없을 경우 $r_{u,i}$는 가상 평가 값인 $S_{u,i}$ 값으로 대체된다.

$$w_{a,u} = \frac{\sum_{i=1}^{m} (r_{a,i} - \overline{r_a}) * (r_{u,i} - \overline{r_u})}{\sigma_a * \sigma_u} \qquad (3)$$

$w_{a,u}$는 사용자 a와 이웃 u간의 유사도 가중치이다. 계산된 유사도를 기반으로 지정된 근접 이웃의 수를 참조하여 사용자가 원하는 대상의 평가 예측치를 식 (4)을 이용하여 계산한다.

$$P_{a,i} = \overline{r_a} + k_{weight} \frac{\sum_{u=1}^{n} (r_{u,i} - \overline{r_u}) * w_{a,u}}{\sum_{u=1}^{n} w_{a,u}} \qquad (4)$$

많은 평가를 한 사용자가 적은 평가를 한 사용자 보다 신뢰도가 높다는 점을 고려하여[15] ,평가를 적게한 사용자와 많이 한 사용자를 차별하기 위하여 식(5)에 중요도 가중치 Kweight를 반영하였다[16].

$$k_{weight} = \frac{I_i}{m} \ : \ \text{if} I_i < m, else \ I_i = 1 \qquad (5)$$

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 시스템에 등재되어 있는 140건의 교육자료(PPT형태)를 이용하여 실험을 하였다. 40명의 초등학교 5학년 학생을 대상으로 학기말에 그동안 공부하였던 항목의 자료를 평가하게 하였다. 취득된 데이터 집합은 1 ~ 5 사이의 점수로 평가된 3,214개의 학습 자료 평점으로 최소 30개 이상의 영화를 평가한 29명의 사용자가 본 3,093개 학습 평가 자료를 대상으로 구성되어 있다. 사용자의 정보로는 성별, 과학성적, 과학관심도, 전체성적,PC능숙도, 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 1에서 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 구성시 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.

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

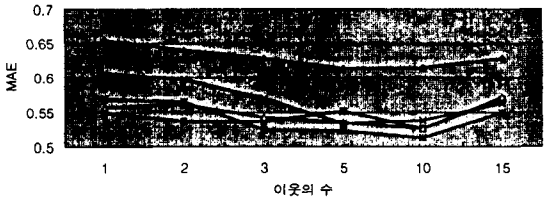

그림 2 사용자 정보를 반영한 실험결과

그림 2는 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. 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 1은 ROC(Receiver Operating Character) 측정에 대한 비교이다[11]. 전체 자료의 평균 값이 3 주변이어서 ROC-3으로 비교하였다.

표 1 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

ROC-3에서는 사용자의 평가 정보중 3, 4, 5의 값은 좋은값(positive)으로, 1, 2의 값은 나쁜값(negative)로 정의한다. Sensitivity는 임의로 선택된 평가값이 좋은값으로 추천될 확률로, 그 값이 1인 경우 완벽한 경우이며, 0.5인 경우 무작위(random)로 판별한다.[11] Specificity는 임의로 선택된 평가값이 나쁜값이 추천되지 않을 확률이다. Accuracy는 전체 실험 자료중 예측이 맞은 경우의 확률이며, Error rate는 전체 실험 자료중 예측이 틀린 경우의 확률이다. 사용자 정보를 반영한 경우 기대했던 바와 같이 정확도가 좋아짐을 확인할 수 있었다. 정보를 적용할 것인가를 선택하는 작업은 문제 영역에 종속적이어서 매우 어렵다. 본 실험에서는 7가지의 사용자 정보중 실험 반영시 영향력이 큰 3개 항목만을 제안 시스템에 반영하였다. 사용자 정보의 선택이나 사용자 정보중 어떤 항목을 제안에 사용할 것인가는 보다 심도 있는 연구가 진행되어야 할 부분이다.

5. 결 론

제안된 추천기법은 협업 여과 추천 기법의 희소성으로 인한 예측의 정확도를 향상시키기 위하여 기존에 주로 이용된 문제 영역의 축소 방식이나, 계산식의 수정이 아닌 희소성의 근본적인 문제점인 빈자리를 인구 통계 정보를 이용하여 채워줌으로써 평가 예측 계산 결과를 향상시킬 수 있는 방법을 제안하였다. 상관관계를 이용하는 유사도 계산식은 비교 대상의 상호성에 대한 자료가 확보되어 있는 경우 보다 정확한 결과를 얻을수 있다. 따라서 상호성이 보장될 수 있도록 가상 평균값을 활용하여 유사도 계산 정확도를 높혀, 결과적으로 예측 결과를 높힐수 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 인구 통계 정보를 보완하면 보다 높은 평가예측의 가능성이 있음을 확인하였으며, 스테레오타입을 이용하는 방식을 도입하기 위한 기초를 마련하였다는데 이 연구의 의미가 있다.

참고문헌

[1] Wattenberg,F. A National Digital Library for Science, Mathematics, Engneering, and Technical Education,D-Lib Magagin, 5,(10 Oct.), 1999.

[2] Wallance, R.,Kuppeman,J.,Krajcik,J., and Soloway,E. "Science on the Web: Students on line in sixth-grade classroom", Journal of the Learning Sciences, 9(1), pp75~104, 2000.

[3] Borgman,D.,Krieger,D., Gallagher, Al,&Bower,J. "Children's use of an interactive science library :Exploratiory research", School Library Media Quaterly, 18, pp108~113, 1990.

[4] Goldberg,D., Nichols,D., Oki,B.M., and Terry,D, "Using Collaborative Filtering to Weave an Information Tapestry", Communications of the ACM, 35(12) pp61~70, 1992

[5] Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, K., and Riedl, J, "GroupLens:Applying Collaborative Filtering to Usenet News", Communications of the ACM, 40(3), pp77~87, 1997.

[6] Shardanand, U.; and Maes, P., "Social information filtering: Algorithms for automating word of mouth", Proceedings of ACM CHI'95, Conference on Human Factors in Computing Systems, pp210~217, 1995.

[7] Basu, C., Hirsh, H., and Cohen, W, "Recommendation as Classification : Using Social and Content-based Information in Recommendation", In Recommender System Workshop, pp11~15, 1998.

[8] Recker,Mimi M., Walker Andrew, Lawless Kimberly, "What do you recommend ? Implementation and analysis of collaborative filtering of Web resources for education",Proceedings of International Conference on Artificial Intelligent,http::/it.usu.edu/~mini/papers /instscience1.doc 2002.

[9] Soboroff,I. ,Nocholas,C., "Combining content and collaborative in text filtering",Proceedings of the IJCAI Workshop on Machine Learning in Information Filtering, pp.86~9, 1999.

[10] Sarwar,B.M., Karypis, G., Konstan, J. A., and Riedl, J., "Application of Dimensionality Reduction in Recommendation System-A Case Study" ACM WebKDD Web Mining for E-Commerce Workshop, http://robotics.stanford.edu/~ronnyk/WEB KDD2000/papers/,2000.

[11] Pazzani,M.J. ,"A Framework for Collaborative, Content-Based and Demographic Filtering", Artificial Intelligent Review, pp394~408, 1999.

[12] Balabanovic, M., and shoham, Y., "Fab : Content-based, collaborative recommendation", Communications of the Association of Computer Machinery 40(30), pp66~72,1997.

[13] Good,N., Schafer, B., Konstan,J., Borchers, A.Sarwar, B., Herlocker, J.,Riedle J., "Combining Collaborative Filtering with Personal Agents for Better Recommendation", Proc. of the AAAI conference. pp439~446,1999.

[14] Sarwar,B.M., G.Karypis, J.A.Kostan, and J.A., & Riedl,J, "Getting to Know you :Learning New User Preferences in Recommender System for Groups of Users",Proc. of the 7th International conference on Intelligent user interfaces, pp.127-134, 2002.

[15] Melville,P., Mooney,R., Nagarajan, R, " Content-Boosted Collaborative Filtering for Improved Recommendations" Proceedings of the eighteenth National Conference on Artificial Intelligence, pp187~192,2002.

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