제 19 회 한국정보처리학회 춘계학술발표대회 논문집 제 10 권 제 1 호 (2003. 5)

신경망 기반 협력적 여과의 성능 향상을 위한 연구

김은주*, 류정우*, 김명원*
*숭실대학교 컴퓨터학과
e-mail : blue7786@bineee.pe.kr

A Study on Collaborative filtering Based on Neural Network for Increment Performance

Eun-Ju Kim*, Joung-Woo Ryu*, Myung-Won Kim*
*Dept. of Computing, Soongsil University

요 약

추천 시스템을 위한 여과 기술에는 협력적 여과, 내용기반 여과 등이 있다. 협력적 여과 방법은 적용이 용이한 반면 희소성 문제와 초기 평가 문제가 있으며, 내용기반 여과는 정보의 질을 구분하는 것이 어려워 효과가 적다는 단점이 있다. 신경망 기반 협력적 여과 방법은 이러한 문제를 해결하고 있지만, 사용자의 수가 많아 지면 모델이 커져 효율성이 떨어지는 문제가 있다. 본 논문에서는 신경망 기반 협력적 여과의 효율성을 높이기 위해 상관도를 고려하는 신경망 기반 협력적 여과를 제안한다. 여기서 상관도란 피어슨 상관계수를 이용하여 구해진 상관계수의 절대값을 의미하며, 상관도가 높다라는 것은 상관계수의 절대값이 1 에 가까운 경우를 말한다. 본 논문에서는 EachMovie 데이터를 이용하여 제안한 방법의 우수함을 보인다.

1. 서론

현재 웹사이트를 이용한 전자상거래 상에서 이루어지고 있는 추천 시스템은 사용자들의 관심이나 구매 기록에 대한 입력을 이용하여 항목을 추천해준다. 추천 시스템의 대표적인 방법에는 협력적 여과(collaborative filtering), 내용기반 여과(content-based filtering) 등이 있다[1,2].

협력적 여과는 목표 사용자와 유사한 선호도를 갖는 다른 사용자들의 선호도를 바탕으로 항목에 대한 목표 사용자의 선호도를 추정하는 방법이다. 이 방법은 적용이 용이한 장점을 가진 반면 사용자들 간의 유사도를 계산하는데 있어 항목간의 중요도 즉, 가중치를 고려하지 못하고 사용자가 선호도를 표시한 항목의 개수가 적을 경우 희소성(sparsity)문제가 발생할 수 이는 단점이 있다. 내용기반 여과는 항목의 다양한 속성 정보를 이용하여 사용자가 선호하는 항목을 추정하는 방법이다. 이 방법은 내용 분석에 있어 텍스트의 단순한 비교를 통해 분석이 이루어지고 정보의 질을 구분 하는 것이 어렵기 때문에 효과가 적은 단점이 있다.[3,4,5].

이러한 협력적 여과 방법과 내용 기반 여과의 단점을 보안하기 위하여 [6]에서는 신경망 기반 협력적 여과를 제안하였다. 이 방법은 신경망을 이용하여 사용자 혹은 항목들 간의 선호 상관관계를 학습시킴으로써 모델을 생성하고 그 모델을 사용하여 선호도를 추정한다. 이 방법은 다양한 정보를 융합함으로써 희소성 문제와 초기 사용자 문제를 해결할 수 있으나, 사용자 혹은 항목 수가 많아지면 입력노드가 증가되어 모델이 커짐으로 효율성이 떨어지는 문제를 가진다.

본 논문에서는 신경망 기반 협력적 여과의 효율성을 높이기 위하여 상관도를 고려하는 방법을 제안한다. 이 방법은 피어슨 상관계수(pearson correlation coefficient)로 구해진 상관도를 이용하여 목표 사용자와 상관 있는 사용자들을 바탕으로 신경망 모델을 생성하고 목표 사용자의 항목에 대한 선호도를 예측한

본 연구는 한국 과학기술부에서 지원하는 뇌신경 정보학 연구사업으로 수행되었음

다.

본 논문의 구성은 다음과 같다. 2 장에서는 기존 최근접 이웃 방법에 의한 협력적 여과 방법과 신경망 기반 협력적 여과 방법에 대해 기술한다. 3 장에서는 상관도를 고려한 신경망 기반 협력적 여과에 대해 기술하고, 4 장에서는 신경망 기반 협력적 여과와 상관도를 고려한 신경망 기반 협력적 여과를 비교 실험을 하며, 마지막으로 5 장에서는 결론을 맺고 향후 연구를 제시한다.

2. 관련 연구

2.1 협력적 여과

협력적 여과는 사용자들 간의 상관관계를 이용하여 관심 있는 정보들을 찾아내고 정보의 선호도를 추정하는 방법이다. 협력적 여과의 대표적인 방법인 최근접 이웃 방법(nearest neighbor method)은 GroupLens 에서 처음 제안 되었다. 이 방법은 가장 가까운 이웃을 찾아 새로운 사용자에 대한 예측 및 분류작업을 하는데 사용되는 메모리 기반(memory-based)협력적 여과 방법이다. 메모리 기반이란 예측을 위해 전체 데이터들을 사용하는 것을 의미한다. 따라서 최근접 이웃 방법은 데이터가 증가할수록 수행속도 저하와 메모리가 증가되는 범위성(scalability) 문제를 가지고 있다.

$$w_{a,u} = \frac{\sum_{i=1}^{m}(r_{a,i} - \bar{r_a})(r_{u,i} - \bar{r_u})}{m(\sigma_a \times \sigma_u)} \qquad \text{식 1}$$

$$p_{a,j} = \bar{r_a} + \frac{\sum_{u=1}^{k}\left[w_{a,u}(r_{u,j} - \bar{r_u})\right]}{\sum_{u=1}^{k} w_{a,u}} \qquad \text{식 2}$$

최근접 이웃 방법은 사용자들 간의 유사도를 피어슨 상관계수(Pearson Correlation Coefficient)를 이용하여 계산한다. 피어슨 상관계수는 두 사용자 간의 선형 상관관계의 정도를 -1 에서 1 의 값으로 나타내며, 1 에 가까워질수록 두 사용자 간의 양의 상관관계가 존재한다는 것을 말한다. 양의 상관관계란 사용자(a)가 선호하는 항목을 사용자(u)도 선호한다는 것을 의미한다. 반대로, -1 에 가까워질수록 사용자(a)가 선호하는 항목을 사용자 (u)는 선호하지 않는다는 음의 상관관계가 존재한다는 것을 의미한다. 또한 상관계수가 0 에 가까우면 선형 상관관계가 적다는 것을 나타낸다. 피어슨 상관계수는 <식 1>을 이용하여 계산되는데 $w_{a,u}$ 는 사용자 a 와 u 의 피어슨 상관계수를 의미하며, $r_{a,i}$ 는 항목(i)에 대한 사용자(a)의 선호도를 나타내고, $\bar{r_a}$, σ_a 는 각각 사용자(a)가 선호도를 표시한 모든 항목에 대한 선호도의 평균과 표준편차를 나타내고 있다. <식 2>은 피어슨 상관계수를 가중치로 하여 사용자(a)의 항목(j)에 대한 선호도 정보를 예측한다. 여기서 k 는 사용자(a)와 유사한 사용자의 수이다[1,3,4].

2.2 내용기반 여과

내용기반 여과(content-based filtering)는 정보 검색(information retrieval)을 기반으로 하고 있으며, 사용자가 관심을 갖는 항목의 내용의 분석을 통해 추천을 하는 방법이다. 이 방법은 추천 시스템에서 사용자가 선호하는 항목 내용의 분석으로 만들어진 사용자 프로파일(user profile)과 문서 내용 사이의 비교를 통하여 추천이 이루어진다.

이러한 내용기반 여과는 내용의 분석에 있어서 단순히 텍스트간의 비교를 통한 분석이 이루어지고 정보의 질을 구분하는 것이 어렵기 때문에 추천의 효과가 적은 단점이 있다[5].

2.3 신경망 기반 협력적 여과

기존 협력적 여과 방법은 적용이 용이한 반면 항목의 종류가 많은 데이터일 경우 사용자가 선호도를 표시한 항목의 개수가 적어 사용자 간의 상관관계가 왜곡되는 희소성(sparsity)문제와 사용자간의 가중치를 고려하지 못하는 문제가 발생할 수 있다.

이러한 기존 추천 방법의 단점을 보완하기 위한 방법으로 [6]에서는 신경망 기반 협력적 여과 방법을 제안하였다. 신경망 기반 협력적 여과 방법은 신경망을 이용하여 사용자 혹은 항목들 간의 선호 상관관계를 학습시킴으로써 모델을 생성하고 그 모델을 사용하여 선호도를 추정하는 방법이다.

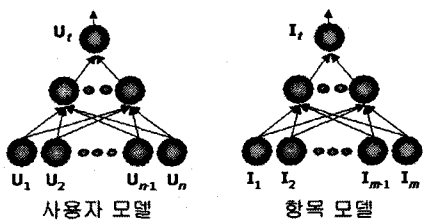

그림 1. 사용자 모델과 항목 모델

신경망 기반 협력적 여과 방법에는 사용자, 항목 모델이 있다. 사용자 모델은 <그림 1>의 왼쪽에서처럼 목표 사용자(U_t)와 다른 사용자 <U_1, U_2, …, U_{n-1}, U_n>들 간의 상관관계를 목표 사용자가 선호했던 항목들을 가지고 학습하여 생성된다. 추천 항목이 주어지면 그 항목에 대한 다른 사용자들의 선호도를 입력으로 하여 추천항목에 대한 목표 사용자의 선호도를 추정한다. 항목 모델의 경우는 사용자 모델과 달리 목표 항목(I_t)와 다른 항목 <I_1, I_2, … , I_{m-1}, I_m>들 간의 선호 상관관계를 목표 항목을 선호했던 사용자들의 정보를 이용하여 학습하고 모델을 생성한다.

또 [6]에서는 희소성문제를 해결하기 위하여 다양한 정보를 융합한 신경망 기반 협력적 여과를 제안하였다. 이 방법은 <그림 2>와 같이 사용자 모델에 내용정보를 항목 모델에 인구통계학적 정보를 융합한다. 기존 방법들이 이질적인 정보를 융합하기 위해 전처리를 수행하거나 또 다른 융합 방법을 고려해야 하는 어려움이 있는 반면 이 방법은 단지 신경망에 입력 노드를 추가함으로써 쉽게 융합할 수 있는 장점이 있다.

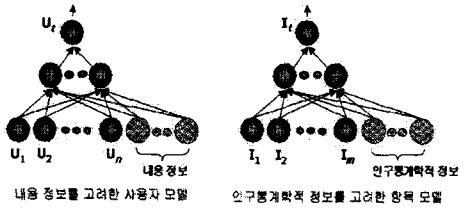

그림 2. 다양한 정보를 융합한
신경망 기반 협력적 여과

3. 상관도를 이용한 신경망 기반 협력적 여과

2.3 절에서와 같이 신경망 기반 협력적 여과 방법은 다양한 정보를 융합함으로써 희소성문제를 해결하고 있지만 사용자 혹은 항목의 수가 많아지면 입력노드가 증가되어 모델이 커짐으로 효율성이 떨어지는 문제를 가진다. 따라서 본 논문에서는 효율성과 추천 성능을 높이기 위하여 상관도를 고려한 신경망 기반 협력적 여과 방법을 제안한다.

제안된 방법은 <그림 3>과 같이 전체 사용자 n 명에서 목표 사용자와 상관이 있는 사용자 즉, 상관도가 높은 순으로 k 명 선택하여 신경망 모델을 생성한다. 본 논문에서 상관도가 높다는 것은 <식 1>에 의해 구해진 상관계수의 절대값이 1 에 가깝다는 것을 의미한다. 즉, 상관도가 0.7 이라는 것은 양의 상관계수인 0.7 과 음의 상관계수인 -0.7 을 모두 고려하는 것을 말한다.

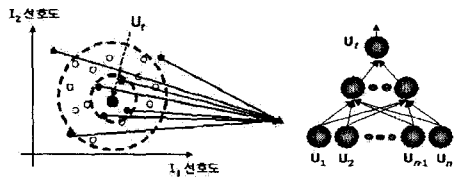

그림 3. 상관도를 고려한 신경망 기반 협력적 여과

특히 <그림 4>와 같이 양의 상관관계를 가지는 경우를 유사도라고 하고 유사도만을 고려하는 신경망 기반 협력적 여과를 유사도를 이용한 신경망 기반 협력적 여과라고 표시한다.

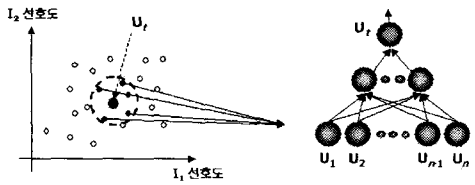

그림 4. 유사도를 고려한 신경망 기반 협력적 여과

4. 실험 및 결과

4.1 실험 방법

본 실험에서는 실험 데이터로 EachMovie[8] 데이터에서 최소 100 편 이상 영화를 본 1000 명의 사용자를 선택하고, 그 중 선호 비율이 40~60%로 편향되지 않으며 음의 상관관계를 10%이상 포함한 사용자 10 명과 항목 10 개를 선택하여 모델을 생성한다. 선호도가 입력되지 않은 경우를 처리하기 위해 <표 1>과 같이 데이터를 정량화하여 이용한다.

표 1. 정량화 방법

선호도	0.0	0.2	0.4	0.6	0.8	1.0	결측값
정량화	-1.0	-0.6	-0.2	0.2	0.6	1.0	0.0

선택된 모델들은 4-fold cross validation 으로 평가된다. 모델을 생성할 때 목표항목을 제외한 모든 속성이 입력노드로 설정되어야 하지만 본 실험에서는 100 개의 입력노드로 한정시킨다 따라서 상관도를 고려하는 하지 않은 경우에는 임의의 100 명의 사용자를 선택하고 상관도를 고려하는 경우에는 상관도가 높은 100 명을 선택하며 유사도를 고려하는 경우에는 유사도가 높은 100 명을 선택하여 모델을 생성한다.

4.2 평가 방법

추천 성능 척도는 accuracy, precision, recall, F-Measure 를 이용한다. 기계학습에서 사용되는 accuracy 는 항목들을 선호, 불호로 얼마나 잘 분류되었는가를 나타내는 척도이다. 정보검색에서 사용되는 표준 척도인 precision 은 추천된 항목들 중에서 사용자가 선호하는 항목들의 비율을 나타내며, recall 은 사용자가 선호하는 항목 수에 대한 추천된 항목 수의 비율을 나타낸다. F-measure 는 Precision 과 recall 의 조화평균으로 두 척도를 동시에 고려한다.

4.3 실험 결과

표 2. 사용자 모델의 실험 결과

	임의 추출	유사도	상관도
accuracy	72.9	74.1	81.8
precision	70.4	70.7	80.2
recall	70.1	71.8	79.6
F-measure	69.2	70.5	79.1

실험의 결과 사용자 모델의 경우 <표 2>에서처럼 임의 추출을 이용한 신경망 기반 협력적 여과방법에 비하여 상관도를 고려한 신경망 기반 협력적 여과방법이 약 9%, 정도의 성능이 향상되었다.

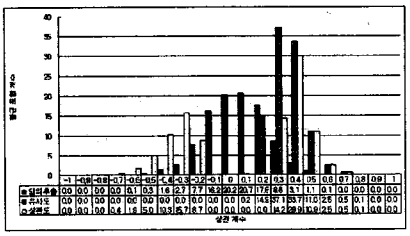

그림 5 사용자 모델의 상관계수 별 분포

<그림 5>는 사용자 모델에 대한 임의적 방법, 유사도, 상관도를 고려하여 추출한 사용자들의 상관계수 분포이다. 그림에서처럼 임의 추출한 경우에는 약한 상관관계를 갖는 0 값 전후에 많이 분포되어 있으며, 유사도를 고려한 경우는 평균적으로 0.22 ~ 0.58 사이에 분포 되어있고, 상관도는 절대값의 평균적으로 절대값이 0.31~0.58 사이에 분포되어 있다.

표 3. 항목 모델의 실험 결과

	임의추출	유사도	상관도
accuracy	71.4	78.4	78.9
precision	66.6	78.2	77.0
recall	66.8	74.0	75.4
F-measure	66.2	75.2	75.7

항목 모델의 실험 결과는 <표 3>에서처럼 임의 추출에 비하여 상관도를 고려한 경우가 7%정도 향상되었다. <그림 6>은 항목 모델의 임의추출과 유사도 상관도 분포를 나타낸 것이다.

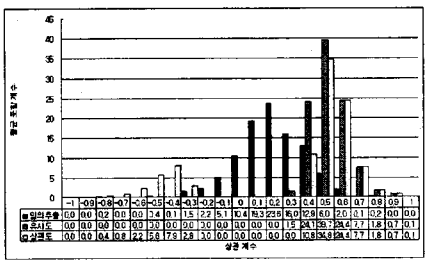

그림 6. 항목모델의 상관계수 별 분포

사용자의 경우 유사도를 이용한 경우보다 상관도를 이용한 경우가 7.7%정도로 성능이 향상되었다. 이것은 항목의 경우 0.5% 향상된 것에 비하여 큰 폭으로 성능이 향상된 것이다. <그림 6>에서처럼 항목에서 유사도 만을 고려하는 경우 평균 상관계수가 0.48 로 비교적 목표 항목과 관련성 있는 항목을 이용하여 모델을 생성하는 것과 달리 <그림 5>에서는 사용자에서 유사도 만을 고려하는 경우 상관계수가 0.22 로 모델 생성에 있어서 관련성 있는 사용자를 이용하기 힘들기 때문이다. 즉, 사용자의 경우와 같이 약한 양의 상관관계를 보이는 경우 음의 상관관계가 성능 향상에 도움이 되어 유사도만을 고려한 경우보다 상관도를 고려한 경우에 성능 향상의 폭이 크다.

5. 결론

본 논문에서는 상관도를 이용한 신경망 기반 협력적 여과를 제안하였다. 이 방법은 기존 협력적 여과 방법의 단점인 사용자 혹은 항목간의 가중치를 고려하지 못하는 문제를 해결한다 또한, 신경망 기반 협력적 여과에서 사용자 혹은 항목의 수가 많아질 경우 신경망의 입력노드가 증가되어 모델이 커짐으로 효율성이 떨어지는 문제를 해결한다.

본 논문에서는 상관도를 구하기 위하여 피어슨 상관계수를 이용한다. 피어슨 상관계수는 두 사용자간의 선형 상관관계만 고려하므로 비 선형 상관관계를 가지는 경우에는 올바른 결과를 나타내지 못한다. 따라서 향후 연구로는 EM(Expectation Maximization), SOM(Self-Organize Map) 등 다른 클러스터링 기법을 적용하여 비교 실험을 할 것이다. 또한 EashMovie 데이터가 아닌 다른 데이터를 사용하여 추천의 성능을 검증할 것이다.

참고문헌

[1] Linden, G.; Smith, B.; York, J., " Amazon.com recommendations: item-to-item collaborative filtering " , Internet Computing, IEEE , Volume: 7 Issue: 1 , Jan/Feb 2003, Page(s): 76 -80, 2003

[2] Andrew I. Schein, Alexandrin Popescul, Lyle H. Ungar, David M. Pennock, "Methods and Metrics for Cold-Start Recommendations", In Proceedings of the 25'th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002).

[3] Kwok-Wai Cheung, James T. Kwok, Martin H. Law and Kwok-Ching Tsui, "Mining customer product ratings for personalized marketing", Decision Support Systems, Volume 35, Issue 2, May 2003, Pages 231-243, 2003

[4] Sarwar, B. M., Karypis, G., Konstan, J. A., and Ried, J. Analysis of Recommendation Algorithms for E-Commerce. In Proceedings of the ACM EC'00 Conference. Minneapolis, MN. pp. 158-167, 2000.

[5] Mira Kwak, Dong-Sub Cho, "Collaborative filtering with automatic rating for recommendation", Industrial Electronics, 2001. Proceedings. ISIE 2001. IEEE International Symposium on , Volume: 1 , 2001

[6] 김종수, 류정우, 도영아, 김명원, 신경망을 이용한 추천시스템의 성능 향상, 한국뇌학회, Vol.1, No.2, pp.223~244, 2001.

[7] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl. "An algorithmic framework for performing collaborative filtering". In Proceedings on the 22nd annual international ACM SIGIR conference on research and development in information retrieval, pages 230 - 237, Berkeley, CA, August 1999.

[8] P. McJones. Eachmovie collaborative filtering data set. http://www.rearchdigital.com/SRC/each movie, DEC Systems Research Center, 1997