클러스터링을 이용한 신경망 기반 협력적 추천

김은주[0] 류정우 김명원
숭실대학교 컴퓨터학과
{blue7786@naver.com[0], mkim@computing.soongsil.ac.kr

A Collaborative Recommendation Based on Neural Networks Using the Clustering

Eun-Ju Kim[0] Joung-Woo Ryu Myung-Won Kim
Dept. of Computing, Soongsil University

요 약

개인화를 위한 협력적 추천의 대표적인 방법인 최근접 이웃 방법은 적용이 쉽지만, 사용자의 선호도 정보가 적을 경우 희소성(sparsity)문제와 사용자 수가 많은 경우 수행 속도가 느려지는 범위성(Scalability)문제 그리고 사용자간의 가중치가 결여되었다는 점에서 추천의 정확성이 떨어진다. 신경망 기반 추천은 자료의 유형에 상관없이 데이터의 처리가 용이하고, 사용자간의 가중치를 학습할 수 있으며, 내용 정보, 인구통계학적 정보 등을 입력 노드에 추가함으로써 희소성 문제를 해결할 수 있으나. 범위성 문제는 존재한다. 따라서 본 논문에서는 최근접 이웃 방법으로 클러스터링 한 유사한 사용자 또는 항목들을 고려한 신경망 기반 추천 방법을 제안하여 범위성 문제를 최소화시킴으로써 추천의 성능을 향상시키고 있다. 제안한 추천 방법의 타당성을 보이기 위해 EachMovie데이터를 이용하여 기존 신경망 추천과 비교 실험하여 성능을 분석한다.

1. 서 론

개인화를 위한 추천 기술은 협력적 추천(**Collaborative Recommendation**), 내용기반 추천(**content-based Recommendation**), 인구 통계적 추천(**Demographic Recommendation**) 등이 있다.[1][2][3] 협력적 추천은 특정 사용자와 유사한 선호도를 갖는 다른 사용자들의 선호도를 바탕으로 항목의 선호도를 추정하는 기술이다. 내용 기반 추천은 항목의 다양한 속성을 이용하여 사용자가 선호한 항목과 비슷한 속성을 갖는 항목을 추천하는 기술이며, 인구통계학적 추천은 사용자의 나이, 성별, 직업 등 인구통계학적인 정보를 바탕으로 항목이나 정보의 선호도를 추정하는 기술이다.

협력적 추천의 대표적인 방법인 최근접 이웃 방법(**Nearest Neighbor Method**)은 적용하기 용이하나, 사용자 또는 항목들 간의 가중치를 고려하지 못함으로써 추천의 정확도가 떨어지고, 희소성(**Sparsity**) 문제와 범위성(Scalability)문제로 추천의 정확성이 떨어진다.

[4]에서 제안한 신경망 기반 추천은 항목들 또는 사용자들 간의 선호 상관관계를 신경망으로 학습시킴으로써 모델을 생성하고 그 모델을 사용하여 선호도를 추정한다. 단순히 내용 정보나, 인구 통계학적 정보를 입력 노드에 추가하는 것으로 협력적 추천 방법과 내용기반 추천 방법 혹은 인구통계학적 추천방법을 통합할 수 있고, 최근접 이웃방법의 희소성 문제를 해결할 수 있다. 그러나 신경망 기반 추천은 전체 데이터에서 단순임의추출(**Simple Random Sampling**)로 사용자를 선택한 후, 그 사용자들의 선호도 정보를 입력 값으로 사용하고 있어 추천 성능에 영향을 미치고 있다.

본 논문에서는 최근접 이웃방법을 이용한 클러스터링으로 좀 더 유사한 사용자들이나 항목들을 추출하고, 그들의 선호도 정보를 이용하여 신경망으로 학습시키므로 신경망 기반 추천의 성능을 향상시키고, 범위성 문제를 해결할 수 있는 방법을 제안한다.

본 연구는 한국과학기술부에서 지원하는 뇌신경 정보학 연구사업으로 수행되었음.

2. 관련 연구

2.1 최근접 이웃방법(Nearest Neighbor Method)

가장 가까운 이웃을 찾아 새로운 사용자에 대한 예측 및 분류 작업을 하는 데 사용되는 방법이다. 예측 및 분류를 위해 먼저 전체 사용자로부터 학습 자료를 생성하는 메모리 기반(**Memory-based**) 협력적 필터링 알고리즘이다. 이 방법은 새로운 사용자에 대하여 가장 가까운 k개의 근접이웃(**K-Nearest Neighbor**)을 선택하여 다수결 원칙 또는 근접정도에 따른 가중치 평균으로 분류 또는 예측 값을 계산하는 방법이다.

두 사용자의 유사도를 측정하는 대표적인 방법에는 벡터 기반 유사도(**Vector-based similarity**), 상관관계 기반 유사도(**Correlation-based Similarity**)등이 있다. 벡터 유사도의 대표적인 방법으로는 코사인 방법(**Cosine measure**)이 있으며, 상관관계 기반 유사도의 대표적인 방법으로 피어슨 상관 계수(**Pearson Correlation Coefficient**)가 있다.

2.1.1 코사인 방법(Cosine measure)

식 (1)은 2개의 항목 i, j를 m차원 공간에서 두 벡터로 표시한 후, 두 항목의 유사도를 두 벡터사이의 코사인 각도를 이용하여 측정한 방법이다.

$$\cos(\vec{i}, \vec{j}) = \frac{\vec{i} \cdot \vec{j}}{|\vec{i}_2| \times |\vec{j}_2|} \quad (1)$$

2.1.2 피어슨 상관 계수(Pearson Correlation Coefficient)

식 (2)은 두개의 항목 i, j에 대하여 피어슨 상관 계수 $corr(i, j)$를 이용하여 유사도를 측정하는 방법이다.

$$corr(i,j) = \frac{\Sigma_{u \in U}(R_{u,i} - \overline{R_i})(R_{u,j} - \overline{R_j})}{\sqrt{\Sigma_{u \in U}(R_{u,i} - \overline{R_i})^2}\sqrt{\Sigma_{u \in U}(R_{u,j} - \overline{R_j})^2}} \quad (2)$$

여기서 $R_{u,i}$는 항목 i에 대한 사용자 u의 선호도를 말하며,

$\overline{R_i}$은 항목 i에 대한 사용자들의 선호도의 평균을 말한다.[2]

2.2 신경망 추천 모델

신경망 추천 모델은 항목이나 사용자간의 가중치를 학습할 수 있고, 연속 수치형, 이진 논리형, 범주형 등의 자료 유형에 상관없이 데이터 처리가 용이하다. 또한 내용, 인구통계학적 정보 및 항목이나 사용자들 간의 상관관계 등 다른 이질적인 유형의 데이터 및 정보를 통합하기가 용이하다.[4]

2.2.1 사용자와 항목 신경망 모델

신경망 추천 모델에서는 항목들 간 또는 사용자들 간의 선호 상관관계를 [그림 1]과 같이 다층 퍼셉트론으로 모델을 생성하고, 그 모델을 사용하여 선호도를 예측한다[4].

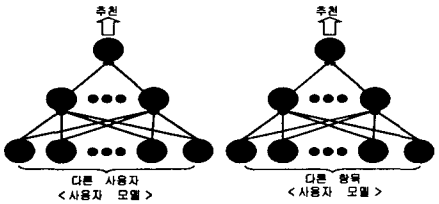

그림 1. 사용자와 항목 신경망 모델

2.2.2 내용 및 인구 통계적 정보를 고려한 신경망 모델

기존의 내용기반 추천, 인구 통계학적 추천이 각각 다른 방법으로 수행되는 데 비하여 [4]에서는 항목에 대한 내용 정보나 사용자의 인구통계학적 정보를 입력 노드에 추가된 [그림.2]와 같은 신경망 모델을 보여주고 있다.

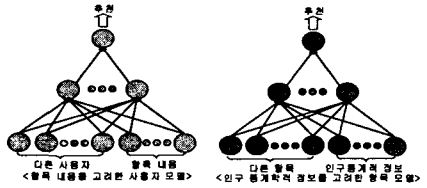

그림 2. 내용 및 인구 통계적 정보를
고려한 신경망 모델

3. 유사도를 고려한 신경망 추천

최근접 이웃 방법은 사용자의 선호 정보가 적을 경우 사용자 간의 유사도가 왜곡 될 수 있는 희소성(**Sparsity**)문제와 사용자의 수가 많을 경우 알고리즘 수행 속도가 느려지는 범위성(**Scalability**) 문제가 존재한다. 신경망 추천의 경우에는 내용 및 인구통계학적 정보를 고려한 신경망 모델을 생성하여 희소성 문제를 감소시키고 있으나, 사용자가 많아지면 신경망 모델이 커지고 학습 속도가 느려지는 범위성 문제를 가지고 있다. [4]에서는 이러한 문제를 최소화하기 위해 [그림 3]에서와 같이 단순임의추출(**Simple Random Sampling**)을 적용하는 그 타당성을 보이고 있다. 즉, 사용자 **A**가 선호도를 표시한 항목에 대하여 임의의 사용자 추출하여 신경망 모델을 생성함으로써 사용자 **A**에 대한 취향을 정확하게 고려할 수는 없다.

본 논문에서는 이러한 범위성 문제에 따른 성능저하를 향상시키기 위해 피어슨 상관 계수로 유사한 사용자를 클러스터링 하는 유사도를 고려한 신경망 추천을 제안한다. [그림 3]과 같이, 최근접 이웃 방법의 피어슨 상관 계수를 이용하여 클러스터링 한 후, 사용자 **A**와 유사한 사용자 즉, 취향이 비슷한 사용자를 추출하고, 그들의 선호도 정보를 신경망 입력으로 이용하여 신경망 모델을 생성하게 된다. 따라서 유사도를 고려한 신경망 추천이 임의의 사용자를 신경망 입력으로 이용하는 기존 신경망 추천보다 사용자의 취향을 보다 정확하게 고려할 수 있다.

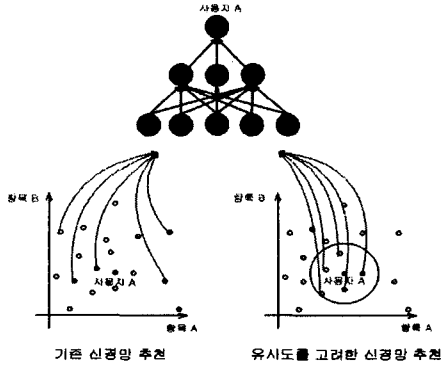

그림 3. 기존 신경망 추천과
유사도를 고려한 신경망 추천

4. 실험

4.1 실험 데이터

실험 데이터는 **EachMovie[5]** 데이터로 **72,912**명의 사용자와 **1,628**개의 영화로 구성되어 있으며, 각 사용자가 본 영화에 대하여 **6**단계의 수치로 선호도를 표현하고 있다.

본 실험에서는 **100**회 이상 선호도를 입력한 **1,000**명의 사용자 중에서 제안한 방법의 타당성을 분석하기 위해 **Like, Dislike**선호도가 편향되지 않은 사용자 **10**명과 항목 **10**개를 선택하여 **4-fold cross validation**으로 각각의 모델을 생성한다.

.

4.2 실험 방법

본 실험에서 사용되는 다층 퍼셉트론 구조는 [4]에서 같이 입력 노드 **100**개, 은닉 노드 **5**개, 출력 노드 **1**개로 설계하고 학습률은 **0.05**로 설정한다

추천의 경우 사용자들 간의 항목에 대한 선호도의 평균적인 경향을 학습하는 것이 중요하므로, 소수의 특정 데이터에 대하여 정확히 학습을 시킬 필요가 없기 때문에 각 모델의 학습은 **MSE(Mean Square Error) 0.04**이하일 때까지만 수행한다. 또한 선호도가 **0.3**이하는 **-1(*dislike*)**, **0.7**이상은 **1(*like*)**로 그리고 중간 선호도와 선호도가 없는 경우는 **0**으로 정량화 하여 입력데이터를 처리하고 있다.

4.3 실험 결과

[표 1]에서처럼 사용자 모델에서는 유사도를 고려한 신경망 추천이 기존 신경망 추천에 비하여 평균 **Accuracy**가 **4.99%** 향상되었다.

표 1. 사용자 모델의 실험 결과

	기존 신경망 추천	유사도를 고려한 신경망 추천
accuracy	77.01	82.00
precision	76.60	80.80
recall	79.93	86.07
F-measure	78.23	83.35

항목 모델의 경우 [표 2]에서처럼 유사도를 고려한 신경망 추천이 기존 신경망 추천에 비하여 **Accuracy**가 평균 **5.37%** 향상되었다.

표 2 항목 모델의 실험 결과

	기존 신경망 추천	유사도를 고려한 신경망 추천
accuracy	73.09	78.46
precision	71.08	78.74
recall	74.34	76.51
F-measure	72.67	77.61

장르를 고려한 사용자 모델의 경우 [표 3]와 같이 유사도를 고려한 신경망 추천이 기존 신경망 추천에 비하여 **Accuracy**가 평균 **6.83%** 향상되었다.

표 3. 장르를 고려한 사용자 모델

	기존 신경망 모델	유사도를 고려한 신경망 모델
accuracy	76.65	83.47
precision	75.86	82.74
recall	80.13	84.56
F-measure	77.94	83.64

4.4. 각 상관계수의 값에 따른 성능 비교 실험

상관 계수의 값에 따른 성능을 비교하고, 추천에 가장 효과적인 상관 계수 값의 임계값을 찾기 위하여 위해서 각 상관 계수의 절대값에 따른 **Accuracy** 변화를 실험하였다. 절대값을 실험에 이용한 이유는 취향이 비슷한 사용자의 선호도뿐만 아니라 취향이 상이한 사용자의 선호도를 고려하여 추천을 하기 위해서이다. 즉, 상관 계수의 절대값이 **0.7**의 모델은 선호도의 상관계수가 **0.7**이상의 값을 가지는 취향이 비슷한 사용자와 **-0.7** 이하의 값을 갖는 취향이 상이한 사용자들을 입력 값으로 이용하여 생성된 모델을 말한다.

표 4. 각 상관 계수 값에 따른 입력노드의 수

상관 계수의 절대값	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
사용자 모델	894	731	527	299	117	27	7	0
항목 모델	727	522	332	176	73	25	9	9

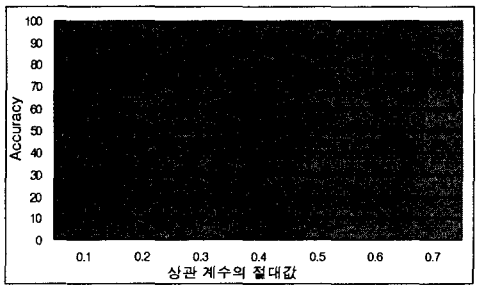

그림 4. 사용자 모델의 상관 계수별 성능 실험

사용자 모델의 상관 계수별 성능 실험에서는 [그림 4]에서처럼 상관 계수의 절대값이 **0.5**일 때 가장 좋은 **Accuracy**를 보였으나 다른 상관계수의 절대값에 비하여 뚜렷이 좋은 결과를 보이지는 못했다. 그러나 [4]의 입력 노드 실험에서 입력 노드의 수가 증가할수록 **Accuracy**가 전반적으로 증가하였으나, 본 논문에서는 입력 노드의 개수가 줄어들어도 성능에는 차이가 없었다. 상관계수의 절대값이 **0.5**일 때의 입력 노드의 수 **117**개, **Accuracy 86.94**로 **0.4**일 때의 입력 노드의 수 **299**개, **Accuracy 84.72**보다 입력 노드의 수는 약 **2/5**감소하였으나, **Accuracy**는 **2.22%** 증가하였다.

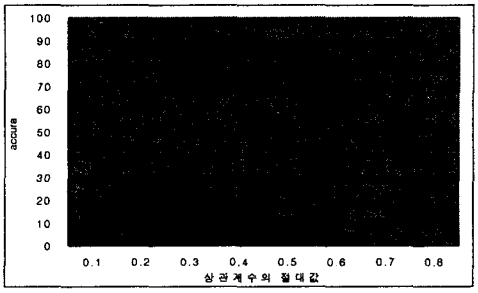

그림 5. 항목 모델의 상관 계수별 성능 실험

항목 모델의 상관 계수별 성능 실험에서도 [그림 5]과 같이 상관 계수의 절대값이 **0.3, 0.4**일 때, 가장 좋은 성능을 보였으나, 다른 상관 계수의 값과는 많은 차이를 보이지는 못했다. 그러나, 사용자 모델의 경우와 마찬가지 상관 계수 절대값이 **0.4**일 때의 입력 노드의 수는 **178**개로, **0.3**일 때 **332**개보다 입력 노드의 개수는 **1/2**줄어들었으나 **Accuracy**는 거의 차이가 없었다.

5. 결론 및 향후 연구계획

본 논문에서는 최근접 이웃 방법을 이용하여 유사한 사용자를 적절히 클러스터링 하여 추출하여 그 들의 선호도를 신경망 입력으로 사용하는 유사도를 고려한 신경망 추천 방법을 제안하고, 기존의 신경망 추천과 비교 실험과 성능에 효과적인 상관 계수의 임계 값을 찾는 실험을 하였다. 이 방법은 기존의 방법에 비하여 사용자 모델, 항목 모델, 장르를 고려한 사용자 모델 모두에서 좋은 성능을 보고, 신경망 입력 노드의 수가 줄어들어도 성능에는 영향을 끼치지 않았다.

향후 연구 계획은 다른 클러스터링 방법을 이용하여 신경망 추천에 가장 효율적인 클러스터링 방법을 찾아보고 ***MovieLens*** 데이터를 이용하여 검증하는 것이다.

6. 참고 문헌

[1] Sarwar, B.M., Karypis, G., Konstan, J.A., and Riedl, J. Item-based Collaborative Filtering Recommender Algorithms. Accepted for publication at the WWW10 Conference. May, 2001

[2] 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.

[3] Breese, J., Heckerman, D. and Kadie, C. Empirical Analysis of predictive Algorithms for Collaborative Filtering. Proceedings of the Fourtheenth Annual Conference on Uncertainty in artificial Intelligence. San Francisco, CA:Morgan Kaufmann, pages 43-52, 1998

[4] 김종수, 류정우, 도영아, 김명원, 신경망을 이용한 추천시스템의 성능 향상. 한국 뇌학회지, Vol 1, No. 2, pp. 233-244, 12월, 2001.

[5] P.McJones. Eachmovie collaborative filtering data set. http://www.research.digital.com/SRC/eachmovie . DEC Systems Research Center, 1997