신경망 기반 추천 모델의 성능향상을 위한 정보의 융합

김호종, 김은주°, 김명원
숭실대학교 컴퓨터학과
usec@hitel.net, blue7786@bineee.pe.kr°, mkim@comp.ssu.ac.kr

Data Fusion for performance Enhancement of Neural Network Based Recommendation Models

Ho Jong Kim, Eun Ju Kim°, Myung Won Kim
Soongsil Univ. Department of Computing

요 약

협력적 추천은 데이터의 범위성, 초기 사용자, 희소성, 회색양의 문제를 안고 있다. 이를 해결하기 위해 기존 연구는 내용기반 추천이나 인구통계학적 추천을 협력적 추천과 통합하려는 연구가 진행되어 왔다. 본 논문에서는 추천 시스템의 성능 향상을 위해 이질적인 데이터의 통합에 효과적인 신경망을 사용하여 다양한 종류의 정보 융합을 제안한다. 신경망을 사용한 추천 모델은 사용자들 또는 항목들 간의 선호관계를 학습할 수 있고, 이질적인 데이터의 통합이 용이한 신경망의 장점을 이용하면 항목들에 대한 내용과 사용자들의 인구통계학적인 정보, 그리고 그 외적인 관련정보를 쉽게 융합할 수 있다. 또한, 데이터 융합을 통하여 희소 데이터 문제와 초기 사용자 문제를 해결할 수 있다.

1. 서 론

개인화 서비스를 위한 추천기술에는 협력적 추천(collaborative recommendation), 내용기반 추천(content-based recommendation), 인구 통계학적 추천(demographic recommendation)이 있다.

협력적 추천은 특정 사용자와 유사한 선호도를 갖는 다른 사용자들의 선호도를 바탕으로 항목에 대한 특정 사용자의 선호도를 추정하는 방법이다. 대표적인 방법으로는 최근접 이웃 방법(k-nearest neighbor)이 있다. 이 방법은 적용이 용이한 반면, 희소성 문제(sparsity problem), 범위성(scalability), 초기 평가(early rater), 회색양(gray sheep)등의 여러 문제를 가지고 있다.[1][2]. 내용기반 추천의 경우 모든 항목들의 내용에 대해 교차하여 적용하는 기술을 사용하기 때문에 협력적 추천에서 발생하는 문제들에 대해 적은 영향을 받는다. 그러나 이 방법은 정보의 질을 구분하는 것이 어려워 추천에 있어 비효율적이다[3]. 내용기반 추천과 협력적 추천의 단점을 보완하기 위하여 [4]에서는 내용기반을 통한 협력적 추천을 제안하였고 초기 사용자 문제를 해결한다. 인구통계학적 추천은 사용자를 표현해 주기 위한 정보로 나이, 성별, 교육정도와 같은 인구통계학적 정보와 항목에 대한 선호정보를 기반으로 추천한다. 그러나 이 방법은 같은 인구통계학적 정보를 가진 사용자라 할지라도 항목에 대한 선호도가 다를 수 있고, 사용자로부터 직접 입력을 받아야 하는 등의 문제점을 가지고 있기 때문에 다른 방법에 비해 추천의 성능이 저하된다. 신경망 추천 모델은 항목들 혹은 사용자들 간의 상관관계를 신경망 가중치로 학습함으로써 모델을 생성하고 추천을 제공하게 된다[4][5]. 게다가 이 방법은 신경망 은닉노드의 개념 형성 기능으로 정확한 선호도 산출이 가능하고 자료 유형에 관계없이 데이터 처리가 용이하다. 즉 연속 수치형, 이진논리형, 범주형 등의 자료 처리가 수월하다.

본 논문에서는 협력적 추천을 위한 신경망 추천 모델에 관련된 추가정보를 융합하여 추천의 성능을 향상 시키는 방법을 제안한다. 추가 정보로는 내용기반 추천에서 사용하는 항목에 대한 내용 정보나, 인구통계학적 추천에서 사용하는 인구통계학적 정보를 사용하고 그 외적인 관련성 있는 정보로써 사용자별 선호 및 불호 장르와 항목의 빈발 여부를 추출하여 사용한다. 이렇게 관련된 추가 정보를 신경망 입력층에서 융합함으로써 초기 사용자에 대한 추천을 제공하고, 희소한 데이터에 대해 추천의 정확성을 높이고자 한다.

본 연구는 한국 과학기술부에서 지원하는 뇌신경 정보학 연구사업으로 수행되었음.

2. 관련연구

2.1 신경망 추천 모델

[3]에서는 항목에 대한 사용자 선호도의 결측치(missing value)를 표현한 뒤 SVD(Singular Value Decomposition)를 사용하여 차원을 축약한 후 이 데이터를 신경망 입력으로 처리하여 학습하는 신경망 추천 방법을 제안하였다. 이것은 항목에 대한 사용자 선호도에 대하여 like는 이진값 "10", dislike는 이진값 "01", 결측치는 "00"으로 표현하고 이를 SVD 적용 후 신경망의 학습데이터로 사용하는 방법이다. 그러나 SVD는 계산시간이 많고, 축약된 데이터가 신경망 입력 데이터로 사용됨으로 인하여 원래의 데이터로 처리했을 때의 장점을 감소시키는 역효과를 낼 수도 있다.

[5]에서는 추천의 정확도를 향상시키기 위하여 신경망 추천 모델을 제안하였다. 신경망 추천 모델은 사용자 신경망 추천 모델과 항목 신경망 추천 모델이 있다. 사용자 신경망 추천 모델은 목표 사용자와 다른 사용자들 간의 항목에 대한 선호상관관계를 이용하여 모델을 생성하고 목표 사용자의 해당 항목에 대한 선호도를 예측한다. 항목 추천 신경망 모델은 사용자 추천 신경망 모델과 달리 목표 항목에 대한 선호도를 다른 항목들의 사용자에 대한 선호도를 이용하여 예측한다.

2.2 신경망 추천 모델의 통합

사용자 신경망 추천 모델은 사용자들 간의 연관성으로 선호도를 예측하고, 항목 신경망 추천 모델은 항목들 간의 연관성으로 선호도를 예측하기 때문에 두 모델은 서로 다른 선호도로 예측 할 수 있다. 그러므로 서로 다른 관점을 통해 예측된 선호도를 통합하여 추천한다면, 한 관점만으로 선호도를 예측하는 단일 신경망 추천 모델이나 기존의 추천 방법보다 유연성과 효율성 모두 향상될 수 있다. [7]에서는 신경망 추천 모델을

통합하는 방법으로 순차적(Sequential), 병렬적(Parallel) 방법과 퍼셉트론(Perceptron)과 다층 퍼셉트론(Multi-Layer Perceptron) 그리고 퍼지 추론과 BKS(Behavior Knowledge Space)를 사용한다. 이 방법들 중 다층 퍼셉트론은 임의의 복잡한 경계선을 근사할 수 있으므로 보다 향상된 통합 성능을 보였다.

3. 신경망을 통한 추가정보의 융합

본 논문에서는 사용자 및 항목 신경망 추천 모델에 추가정보를 사용하여 추천 성능을 높이는 방법을 제안한다. 이 방법은 [7]에서 제안된 방법이 서로 다른 선호도를 예측하는 두 개의 모델을 통합하는 모델 융합(model fusion)방법인 것이 비하여, 사용자 혹은 항목에 추가정보 즉 이질적인 데이터를 융합하는 속성 융합(feature fusion)방법이다. <그림.1> 왼쪽 그림과 같이 사용자 신경망 추천 모델에 항목의 내용 정보를 추가정보로 사용하여 내용기반 추천과 협력적 추천을 병합한 효과를 갖는 모델이 생성되고, 항목 신경망 추천 모델에 인구통계학적 정보를 추가정보로 사용하여 인구통계학적 추천과 협력적 추천을 병합한 효과를 갖는 모델이 생성된다. 추가정보는 신경망 입력 노드에 추가정보를 입력할 수 있을 만큼의 신경망 노드를 생성하여 은닉층의 노드들과 완전연결(fully connected)시킴으로써 정보의 융합이 가능해진다.

그러나 [5]에서는 사용자 신경망 추천 모델에 항목에 대한 추가정보로 장르정보를 사용하여 성능을 향상시킨 반면, 항목 신경망 추천 모델에서는 사용자에 대한 추가정보로 인구통계학적 정보를 사용하였으나 효과적인 추천을 하지 못하였다. 이는 사용자의 인구통계학적 정보가 추천에 도움을 주지 못하는 정보이기 때문이다. 예를 들어 나이도 같고, 성별, 직업도 같은 학생 2명의 취향이 다를 수 있기 때문이다.

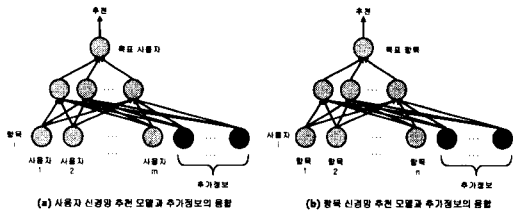

그림 1. 신경망 추천 모델과 추가정보의 융합

따라서 본 논문에서는 인구통계학적 정보보다는 사용자와 항목에 대해 관련성이 높은 정보를 사용하기 위하여 사용자별 선호 및 불호 장르를 추출하여 항목 신경망 추천 모델에 적용한다. 또한 사용자 신경망 추천 모델에 대해서는 항목 빈발 여부를 추가정보로 사용한다.

4. 실험 및 분석

4.1 실험 데이터

실험 데이터로는 Movielens[8] 데이터를 사용한다. Movielens는 총 1682편의 영화에 대해 943명의 사용자가 영화별 선호도를 입력한 데이터로 적어도 20편 이상의 영화에 대해 선호도를 입력하고 인구통계학적 정보를 모두 가진 사용자들의 데이터만을 추출한 데이터이다. 전체 선호도 입력 횟수는 100,000번이고, 영화에 대한 장르는 모두 19가지, 사용자에 대한 인구통계학적 정보는 성별, 나이, 직업이 있으며 직업의 종류는 21가지이다. 선호도의 입력 범위는 1~5로 5단계로 표현되며, 숫자가 커질수록 선호도가 높아지도록 표현되어있다.

사용자 신경망 추천 모델과 항목 신경망 추천 모델은 사용자 또는 항목별 선호도 입력 횟수에 따라 구분하여 30회 미만, 50회 미만, 100회 이상, 150회 이상으로 구분하여 조건에 맞는 데이터를 추출한다.

본 실험에서는 제안한 방법의 타당성을 분석하기 위해 like 비율이 35%~65%사이로 선호도가 편중되지 않고, 선호도 입력 횟수를 100회 이상인 사용자와 항목 각각 30개를 선정하고 4-fold cross validation방법을 이용하여 모델을 생성한다.

4.2 실험 방법

기본 신경망 추천 모델은 입력층 노드는 100개, 은닉층 노드 5개, 그리고 출력층 노드는 1개로 구성되며 가중치는 완전연결로 구성된 다층 퍼셉트론을 사용한다. 신경망 가중치는 -0.5~0.5로 초기화하고 학습률은 0.05로 지정한다. 학습에 대한 종료조건으로는 MSE(Mean Square Error)가 0.04가 될 때까지 학습한다. 일반적인 신경망 학습은 MSE가 최소화 되도록 학습하지만 소수의 특정 데이터에 대해서 정확히 학습시킬 필요는 없다. 게다가 0에 가깝게 충분히 수렴시킬 경우 과다학습(overfitting)의 문제가 발생한다. 실험 시 입력 데이터는 평가에 도움이 안 되는 중간 선호도와 선호도가 없는 경우는 0.0으로 선호 정보(선호도 4이상)는 1.0으로 불호 정보(선호도 2이하)는 -1.0으로 정량화 하여 사용한다.

본 논문에서의 평가척도는 주로 기계학습에서 많이 사용하는 accuracy를 사용한다. accuracy는 신경망 추천 모델에 의해 분류된 항목들이 얼마나 옳게 분류되었는가를 나타내는 척도로써 사용되며 식 (1)과 같다.

$$accuracy = \frac{\left(\begin{array}{c}\text{추천된 항목 중}\\ \text{선호하는}\\ \text{항목 개수}\end{array}\right) + \left(\begin{array}{c}\text{추천되지 않은}\\ \text{항목 중 불호하는}\\ \text{항목 개수}\end{array}\right)}{\text{전체 분류 항목 개수}} \quad (1)$$

4.2.1 추가정보 효과 실험 및 희소 데이터 실험

본 논문에서는 추가정보로 항목에 대한 장르정보와 사용자에 대한 인구통계학적 정보를 사용하며, 그 외적인 관련된 정보로 사용자의 선호 및 불호 장르와 사용자의 선호도 입력 횟수에 기반 하여 항목에 대한 빈발 여부를 추출하여 사용한다.

추가정보를 신경망 추천 모델의 입력층에 새로운 노드를 추가하여 입력함으로써 추가정보의 융합을 수행한다. 즉, 장르정보의 경우 19개의 노드를 추가하며 인구통계학적 정보는 성별의 경우 2개의 노드를, 나이의 경우 세분화 된 나이를 총 7단계로 변환하여 7개의 노드를 추가한다. 직업의 경우도 21개의 노드를 추가하여 사용한다. 빈발 항목 여부에 대한 정보는 1개의 노드만을 사용하여 빈발하면 1, 빈발하지 않으면 0을 입력한다. 마지막으로 선호 및 불호 장르에 대한 정보는 선호 장르 정보만 사용할 경우 사용자 마다 선호하는 2개의 장르를 입력하고, 선불호 장르를 모두 고려하는 경우에는 각각 1개의 장르를 입력하게 되어 총 38개의 노드를 추가하여 사용한다.

학습 및 테스터 데이터 추출 시 사용자의 선호도 입력 횟수에 제한을 두어 추출함으로써 희소 데이터를 생성한다. 본 논문에서는 30회, 50회 미만 선호도 입력에 대해 100회, 150회 이상 선호도 입력보다 희소한 데이터로 간주하여 비교 실험한다.

4.2.2 초기 사용자 실험

초기 사용자 실험에서 학습 데이터 추출은 희소 데이터 실험과 같은 방식으로 데이터를 추출하여 학습하지만 검증 데이터의 경우 초기 사용자 데이터를 만들기 위해서 선호도 입력 횟수가 5회 미만이 될 때까지 선호도를 임의로 제거한다. 그리고 본 논문에서는 이렇게 생성된 검증 데이터를 초기 사용자에 대한 데이터로 간주하고 실험한다.

4.4 실험 결과

4.4.1 추가정보 효과 실험 및 희소 데이터 실험

<표.1>에서와 같이 항목 신경망 추천 모델에 추가정보를 사용하는 경우 7.5%의 성능 향상을 보인다. 추가정보로는 선호와 불호 장르를 동시에 융합한 것이 가장 좋은 성능을 나타내었다. 그러나 인구통계학적 데이터를 추가정보로 사용하였을 때, 선호도의 분포가 유사하여 선호도 예측에는 좋지 않은 추가정보로 분석되었다. 또 추가 정보를 사용하지 않은 30회 미만의 결과와 150회 이상의 결과를 비교하면 약 7.5%의 차이를 보이지만, 선호 장르와 불호 장르를 함께 추가 정보로 사용한 30회 미만의 결과가 150회 이상의 결과보다 약 4.5%의 차이를 보이고 있다.

표 1. 항목 신경망 추천 모델

	없음	인구통계	선호장르	선불호장르
30회 미만	55.2	54.2	57.3	61.7
50회 미만	56.3	57.6	60.0	63.7
100회 이상	62.3	62.8	64.1	65.6
150회 이상	52.7	62.3	63.6	66.2

<표.2>의 사용자 신경망 추천 모델의 결과는 추가정보의 사용으로 30회 미만에서 추가정보를 사용하지 않을 경우와 장르 + 항목 빈발 여부의 경우 약 5.3%, 50회 미만에서는 약 4%정도의 차이를 보인다. 또한 추가정보를 사용하지 않은 30회 미만과 150회 이상의 결과는 약 13%의 차이를 보이나, 항목 빈발 유무를 추가정보로 사용할 경우 약 8.5%의 차이를 보인다.

표 2. 사용자 신경망 추천 모델

	없음	장르	항목 빈발여부	장르+ 빈발여부
30회 미만	55.2	54.2	57.3	61.7
50회 미만	56.3	57.6	60.0	63.7
100회 이상	62.3	62.8	64.1	65.6
150회 이상	52.7	62.3	63.6	66.2

즉, 데이터가 희소해 질수록 추가정보를 사용한 경우가 그렇지 않을 때 보다 인식률의 저하가 둔화되고, 추천 성능이 향상됨을 알 수 있다.

4.4.2 초기 사용자 실험

표 3. 항목 신경망 추천 모델

	30회 미만	50회 미만	100회 이상	150회 이상
없음	55.3	55.6	56.6	56.3
인구통계	54.2	57.6	62.7	62.3
선호장르	57.0	59.7	55.7	55.8
선불호장르	61.1	61.1	56.8	56.5

항목 신경망 추천 모델의 초기 사용자 실험에서는 <표.3>에서와 같이 추가 정보를 사용하지 않을 때 보다 30회, 50회 미만에서는 선호 및 불호 장르 정보를 사용할 때가 약 6% 높고, 100회와 150회 이상에서는 인구통계학적인 정보를 사용하는 것이 약 6% 정도 높은 성능을 나타낸다. 선호도 입력 횟수에 따라 성능의 차이가 있지만 추가 정보를 사용하지 않을 때의 경우보다는 대체로 높은 성능을 나타내고 있다.

표 4. 항목 신경망 추천 모델

	30회 미만	50회 미만	100회 이상	150회 이상
없음	56.8	60.7	65.4	65.0
장르	59.4	62.3	67.0	66.6
항목빈발 여부	57.1	60.1	66.1	65.6
장르+빈발여부	59.5	62.3	67.0	66.8

<표.4>에서와 같이 사용자 신경망 추천 모델에서는 선호도 입력 횟수가 적을수록 초기사용자에 대한 추가정보의 효과가 나타나고 있다. 100회, 150회 이상의 데이터에서는 약 1.5%, 30회와 50회 미만에서는 약 2.5%의 차이를 보인다.

5. 결론 및 향후연구

본 논문에서는 신경망 추천 모델에 관련된 추가 정보를 융합함으로써 기존 협력적 추천 기법의 문제점인 희소 데이터 문제와 초기 사용자 문제를 해결하고자 하였다. 신경망 추천 모델에 관련된 추가 정보의 융합은 항목들 간의 가중치나 사용자들 간의 가중치를 학습할 수 있는 신경망 추천 모델의 장점을 그대로 가지며, 또한 자료 유형에 상관없이 처리가 용이한 신경망의 장점을 이용하면 기존의 내용기반 추천이나 인구통계학적 추천과 협력적 추천을 결합한 형태보다 신경망 추천 모델에 추가정보를 융합하는 처리가 용이하다.

실험결과 제안한 방법은 협력적 추천이 데이터가 적을 경우, 즉 협력적 추천에 기반이 되는 평가된 정보를 충분히 제공받지 못하는 경우 효과적이지 못한데 반해 추가정보를 사용함으로써 희소 데이터 문제와 초기 사용자 문제를 해결할 수 있음을 보였다.

6. 참고문헌

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

[2] M.Claypool, A.Gokhale, T.Miranda, P.Murnikov, D.Netes, and M.Sartin, Combining Content-Based and Collaborative Filters in an Online Newspaper. In Proceedings of the ACM SIGIR'99 Workshop on Recommender Systems: Algorithms and Evaluation. University of California, Berkeley, Aug, 1999.

[3] D. Billsus and M. J. Pazzani. Learning Collaborative Information Filters. In Proceedings of the Fifteenth International Conference on Machine Learning, pages 46-54, Madison, WI, Morgan Kaufman, 1999.

[4] Michael J. Pazzani. A Framework for Collaborative, Content-Based and Demographic Filtering. Artificial Intelligence Review 13(5-6): pages 393-408, 1999.

[5] 김종수, 류정우, 도영아, 김명원, 신경망을 이용한 추천시스템, 한국 뇌학회 학술대회, p110-111, 6월, 2001.

[6] P. Resnick, N. Iacovou, M. Sushak, P. Bergstorm, and J. Riedl, GroupLens: An Open Architecture for Collaborative Filtering of NetNews. In Proceedings of Computer Supported Cooperative Work Conference(CSCW), pages 175-186, ACM SIG Computer Supported Cooperative Work, 1994.

[7] 도영아, 김종수, 류정우, 김명원, 협력적 추천을 위한 효율적인 통합 방법, 한국정보과학회 추계발표 논문집(II), pp130-132, 2001.

[8] MovieLens "http://movielens.umn.edu/"