신경망에 기반한 개인화 기술

김종수[0] 도영아 류정우 김명원
숭실대학교 컴퓨터학과
kjongsu@orgio.net mkim@computing.soongsil.ac.kr

A Personalization Technology Based on Neural Networks

Jong_Su Kim[0] Young-A Do Joung-Wo Ryu Myung-Won Kim
Dept. of Computing, Soongsil University

요 약

현 인터넷상에서 취향에 맞는 항목(상품) 정보를 사용자에게 추천해 주는 개인화 기술은 대부분 특정 사용자와 유사한 선호도를 갖는 다른 사용자들의 특정 항목에 대한 선호도를 바탕으로 항목의 선호도를 추정하는 협력적 추천 기술을 적용하고 있다. 그 중 최근접 이웃 방법은 적용하기가 용이한 반면 항목간의 가중치를 고려하지 못함으로써 추천의 정확도가 크게 떨어지는 문제점이 있다. 연관규칙 방법은 다른 항목에 대한 선호도 자료로부터 데이터 마이닝 기법을 적용하여 항목 선호에 대한 연관규칙을 추출하고 그 규칙을 사용하여 어떤 항목의 선호도를 추정한다. 따라서 항목들 간의 중요도가 연관규칙의 지지도나 신뢰도 등으로 나타난다고 할 수 있으나, 단순히 항목들간의 연관관계 즉 표면적인 연관관계에 의하여 선호도를 결정함으로써 항목들간의 어떤 내용적인 공통성 또는 어떤 상위개념에 의한 선호도가 고려되지 않음으로써 역시 정확도가 떨어지는 문제점이 있다. 본 논문에서는 추천의 정확도를 향상시키기 위한 신경망 추천 방법에 대해 분석하고, 내용기반 추천과 협력적 추천을 병합한 신경망 추천 방법을 제안한다. 또한, 다른 협력적 추천 방법과의 비교를 통하여 본 추천 방법의 장점과 성능의 우수함을 보인다.

1. 서 론

현재 웹사이트 상에서 이루어지고 있는 추천 기술은 크게 협력적 추천(Collaborative filtering), 내용기반 추천(Content-based filtering), 인구통계학적 추천(Demographic filtering)을 들 수 있다.[1] 협력적 추천은 특정 사용자와 유사한 선호도(rating)를 갖는 다른 사용자들의 특정 항목에 대한 선호도를 바탕으로 항목의 선호도를 추정하는 기술이며, 내용기반 추천은 사용자에게 있어 이전에 구매한 항목과 비슷한 속성을 갖는 항목은 사용자 취향에 맞을 가능성이 높다고 보고 선호도가 표시된 항목들의 속성 정보를 이용하여 추천하는 기술이다. 이에 비해 인구통계학적 추천은 사용자의 나이, 성별, 생활수준 등과 같은 인구통계학적 정보를 바탕으로 항목이나 정보의 선호도를 추정한다. 기존의 협력적 추천 중에서 대표적인 방법으로는 최근접 이웃 방법(Nearest Neighbor Method)과 연관규칙 방법(Association Rule Method)등이 있다.[2][3][4][5] 최근접 이웃 방법은 적용하기가 용이한 반면 사용자 또는 항목들 간의 가중치를 고려하지 못함으로써 추천의 정확도가 떨어지는 단점이 있다. 반면, 연관규칙 방법은 다른 항목에 대한 선호도 자료로부터 데이터 마이닝 기법을 적용하여 항목 선호에 대한 연관규칙을 추출하고 그 규칙을 사용하여 어떤 항목의 선호도를 추정한다. 따라서 항목들 간의 중요도가 연관규칙의 지지도(support)나 신뢰도(confidence) 등으로 나타난다고 할 수 있으나 단순히 항목들간의 연관관계 즉 표면적인 연관관계에 의하여 선호도를 결정함으로써 항목들간의 내용적인 공통성 또는 상위 개념에 의한 선호도가 고려되지 않음으로써 역시 정확도가 떨어지는 문제점이 있다. 연관규칙은 그 특성상 선호도가 연속적인 값으로 표시될 경우 이를 이진 값으로 변환해야 하는 문제점이 있다. 또 빈도수가 적은 선호도 항목일 경우 그 항목에 대한 규칙을 생성할 수 없다.

본 논문에서 제안한 추천 방법은 신경망을 이용하는 것으로 항목들 또는 사용자들 간의 선호 상관관계를 신경망으로 학습시킴으로써 모델을 생성하고 그 모델을 사용하여 선호도를 추정한다. 이 방법은 기존의 추천 방법이 가지고 있는 문제점들을 해결할 수 있으며 특히 다음과 같은 장점을 갖는다. 첫째, 항목이나 사용자간의 가중치를 학습할 수 있으므로 보다 정확한 선호도 계산이 가능할 뿐 아니라 신경망 중간노드의 개념 형성 기능으로 정확한 선호도 산출이 가능하다. 둘째, 연속 수치형, 이진 논리형, 범주형 등의 자료 유형에 상관없이 데이터의 처리가 용이하다. 셋째, 다른 이질적(내용, 인구통계학적 정보 등)인 종류의 데이터 및 정보를 통합하기 용이하다. 한편, 기존 방법에서는 내용기반 추천, 인구통계학적 추천이 각각 다른 방법으로 수행되는 데 비하여 신경망 모델에서는 항목에 대한 내용 정보나 사용자의 인구통계학적 정보를 단순히 노드에 추가하여 신경망을 학습시킴으로써 이들 방법들을 용이하게 통합할 수 있는 장점이 있다.

본 연구는 뇌 연구 개발사업
(과제번호 : 98-J04-01-01-A-04) 의 지원을 받았음.

2. 관련연구

2 .1 최근접 이웃 방법(Nearest Neighbor Method)

가장 가까운 이웃을 찾아 새로운 사용자에 대한 예측 및 분류작업을 하는데 사용되는 방법이다. 이 방법은 새로운 사용자에 대하여 전체 고객 자료로부터 가장 가까운 k개의 근접이웃(K-Nearest Neighbor)을 선택하여 다수결 원칙 또는 근접정도에 따른 가중치

평균으로 분류 또는 예측 값을 계산하는 방법이다.[6] 두 사용자의 유사성은 유클리드 거리(Euclidean distance), 코사인 유사도(cosine similarity), 상관관계(correlation)등과 같이 여러 가지 척도로써 나타낼 수 있다. 이와 같은 척도를 사용하여 유사한 사용자 또는 항목을 측정하여 추천하는 방법은 항목의 종류가 많은 데이터일 경우 희소성(sparsity)문제가 발생한다. 희소성 문제란 사용자가 선호도를 표시한 항목의 개수가 적을 경우 사용자간의 유사성이 왜곡될 수 있는 문제를 말한다. 또한, 사용자 수가 많을 경우 알고리즘 수행속도가 현저히 느려지는 범위성(scalability)문제도 고려해야 한다[2][7].

2 .2 연관규칙 방법(Association Rule Method)

연관규칙이란 어떤 사건이 일어나면 다른 사건이 일어난다는 관련성을 의미한다. 이러한 연관규칙이 최근 추천 시스템에서 많이 적용되고 있다. [8]에서는 수집된 사용자의 웹 페이지 네비게이션(navigation) 자료로부터 연관규칙를 추출하여 사용자에게 웹 페이지를 추천하는 시스템을 개발하였다. [5]에서는 최소 신뢰도에 대한 사용자가 원하는 규칙의 개수가 생성될 수 있게 최소 지지도를 자동적으로 조정할 수 있을 뿐만 아니라 미리 선택된 한 개의 항목만이 규칙의 결론부에 나타나도록 알고리즘을 확장하였다. [7]에서는 [5]에서 제안한 알고리즘을 이용한 협력적 추천 방법을 제안하였다.

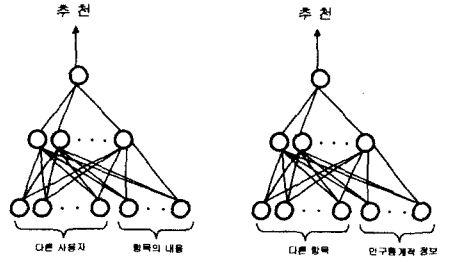
3. 신경망 추천 모델

3.1 사용자와 항목 신경망 모델

신경망 모델은 사용자 신경망 모델과 항목 신경망 모델 두 가지를 사용한다.[1][2][3] 사용자 신경망 모델은 사용자들간의 연관성을 찾아 새로운 항목을 추천하고, 항목 신경망 모델은 서로 다른 사용자들이 평가한 항목간의 연관성을 찾아 새로운 항목을 추천해 준다.

3.2 내용 및 인구통계적 정보를 고려한 신경망 모델

<그림 1>과 같이 내용 및 인구통계학적 정보를 고려하여 신경망을 구현하면 초기 사용자의 경우 협력적 추천에서 추천을 받을 수 없는 문제점을 해결 할 수 있으며, 평가 정보가 적은 사용자가 상품을 추천 받을 경우에도 어느 정도 성능을 향상시킬 수 있다.



<항목 내용을 고려한 신경망 모델> <인구통계적 정보를 고려한 신경망 모델>

그림 1 내용 및 인구통계적 정보를 고려한 신경망 모델

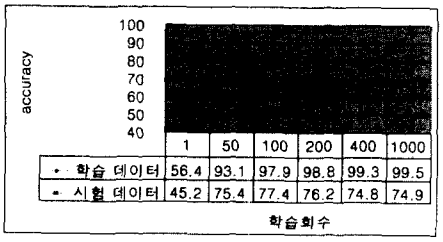
4. 실험 및 결과

4.1 실험 데이터

실험 데이터는 EachMovie[9] 데이터로써 72,916명의 사용자와 1,628개의 영화로 구성되어 있으며 각 고객이 본 영화에 대해서 평가한 선호도(rating)는 0, 0.2, 0.4, 0.6, 0.8, 1.0의 6 단계 수치로 표현되어 있다. 본 실험에서는 최소 100회 이상 선호도를 입력한 사용자 1,000명을 선택하였다. 각 모델에 대해 like 빈도(ratio)는 0.6 ~ 0.4 이고, 학습 데이터(learning data)와 테스트 데이터(test data)는 각각 3:1로 나누었으며, 입력노드의 수는 100개로 한정하여 실험하였다.

4.2 학습 파라미터가 미치는 영향 분석

학습률과 학습회수(learning epoch)는 과다학습(overfitting)과 수렴 속도에 밀접하게 연관되어 있다. 실험 결과 학습률은 0.05에서 가장 적절하였다. 신경망에서의 학습회수는 MSE(Means Square Error)가 0에 충분히 수렴할 때까지 학습을 시키는 것이 보통이다. 그러나 협력적 추천의 경우 사용자들 간의 항목에 대한 선호도의 평균적인 경향을 학습하는 것이 중요하며 소수의 특정 데이터에 대하여 정확히 학습시킬 필요는 없다. 10개의 사용자 모델에 대해 MSE가 0에 수렴할 때까지 학습을 시킬 경우 <그림 2>과 같이 과다학습 문제가 발생한다. 과다학습은 학습회수를 줄이거나 중간층 노드 개수를 적게 하여 모델이 일반적 경향을 나타내도록 생성하여 해결할 수 있으며, 이는 신경망이 갖는 학습속도의 문제점을 보완할 수 있다.



	1	50	100	200	400	1000
학습 데이터	56.4	93.1	97.9	98.8	99.3	99.5
시험 데이터	45.2	75.4	77.4	76.2	74.8	74.9

그림 2 과다학습 현상

4.3 선호도 정량화에 따른 결과

신경망의 장점중 하나인 수치데이터 처리가 용이하다는 점을 이용하여 실제 선호도를 Billsus and Pazzani[10]와는 달리 <표1>과 같이 정량화(quantization)하여 10개의 사용자 모델에 대해 실험을 한 결과 정량화3의 경우 accuracy가 가장 높았다. 즉, 평가에 도움이 안 되는 중간 선호도를 "0"으로 처리함으로써 accuracy를 높일 수 있다.

표 1 선호도 정량화

	선호도가 주어진 경우						선호도가 없는 경우	평균 accuracy(%)
원래 선호도	0	0.2	0.4	0.6	0.8	1	표기 않음	
정량화1	-1	-0.6	-0.2	0.2	0.6	1	0	66.6
정량화2	-1			1			0	62.3
정량화3	-1		0		1		0	77.0

4.4 입력노드 개수에 따른 결과

10개의 항목 모델에 대해 입력 노드 개수를 점차 늘려 가면서 성능의 변화를 분석한 결과 <그림 3>에서 와 같이 accuracy가 증가함을 알 수 있다.

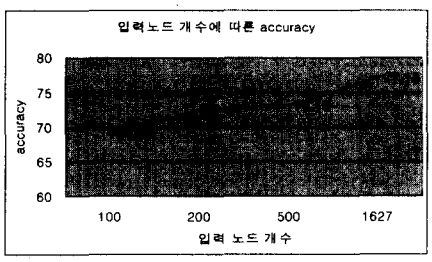


그림 3 입력노드 개수에 따른 accuracy 변화

4.5 사용자 와 항목 모델 결과

13개의 모델을 가지고 사용자 모델, 항목 모델, 장르를 고려한 사용자 모델의 성능을 비교한 결과 <표 2>에서와 같이 장르를 고려한 사용자 모델이 가장 좋았다.

표 2 사용자와 항목 모델 실험 결과

	사용자 모델	항목 모델	장르를 고려한 사용자 모델
accuracy	77.0	71.7	79.7
precision	82.5	70.8	83.9
recall	83.2	73.3	86.3
F-measure	82.8	72.0	85.1

4.6 기존의 추천 방법과의 비교

Lin, Alvarez and Ruiz[5]는 연관규칙 방법을 이용하여 Billsus and Pazzani[10]와 같은 조건하에서 실험을 하였다. 그 결과 accuracy는 67 ~ 69로 서로 비슷함을 보였다. 따라서, 우리는 연관규칙 추천과의 비교를 위해 [5]와 같은 조건인 최소 100회 이상 선호도를 입력한 사용자 1,000명을 학습 데이터로 사용하고, 70,000 이상인 사용자 ID에서 최소 100회 이상 선호도를 입력한 사용자 100명을 선택하여 테스트 데이터로 사용하여 실험하였다. 30개의 모델에 대해 우리가 제안한 추천 방법과 기존의 연관규칙 추천 방법[5][10]을 비교한 결과 <표 3>와 같이 본 논문에서 제안한 신경망 추천이 기존의 연관규칙 추천보다 더 우수함을 알 수 있다. 본 신경망 추천에서는 like와 dislike를 동시에 고려하였으며, 중간노드에서 항목간의 가중치를 학습하기 때문이다.

표 3 기존의 연관규칙 추천과 신경망 추천과의 비교

	연관규칙 추천		신경망 추천		
	사용자 모델	항목 모델	사용자 모델	항목 모델	장르를 고려한 사용자 모델
accuracy	72.0	61.1	81.6	77.5	81.4
precision	75.1	75.4	77.4	76.3	78.0
recall	58.4	22.6	69.6	73.0	65.7
F-measure	65.7	34.8	73.3	74.6	71.3

5. 결론 및 향후 연구

본 논문에서는 추천의 정확도를 향상시키기 위해 신경망 추천 방법에 대해서 분석하고, 내용기반 추천과 협력적 추천을 병합한 신경망 추천 방법을 제안하였다. 실험 결과 본 논문에서 제시한 신경망 추천 방법은 중간노드에서 항목간의 가중치를 고려하며, 자료 유형에 상관없이 데이터 처리가 용이할 뿐만 아니라 항목에 대한 내용정보나 사용자의 인구통계학적 정보를 고려하여 신경망을 학습시킴으로써 기존 추천 방법들 보다 좋은 성능을 보였다.

현재 추천에서 대용량 데이터에서의 범위성이 큰 문제가 되고 있다. 본 논문에서도 단순히 입력 노드개수를 랜덤하게 100개로 한정하여 실험을 하였으나 향후연구로는 이러한 범위성 문제를 해결하기 위해 클러스터링 개념을 도입하고자 한다. 또한, 실험 결과를 토대로 실제 적용 가능한 신경망 추천 시스템을 구현하여 인터넷상에서 개인화 서비스를 행할 수 있도록 하고자 한다.

6. 참고 문헌

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

[2] 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)

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

[4] Weiyan Lin, Sergio A. Alvarez, Carolina Ruiz. Collaborative Recommendation via Adaptive Association Rule Mining. International Workshop on Web Mining for E-Commerce(WEBKDD2000). held in conjunction with the Sixth International Conference on Knowledge Discovery and Dat Mining(KDD2000)

[5] W. Lin, C. Ruiz, and S. A. Alvarez. A new adaptive-support algorithm for association rule Mining. Technical Report WPI-CS-TR-00-13, Department of Computer Science, Worcester Polytechnic Institute, May (2000)

[6] Web_Collaborative Filtering: Recommending Music by Crawling The Web. SOURCE Computer Networks-The International Journal of Computer & Telecommunications Networking, V.33 N.1-6, pages 685-698 (2000)

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

[8] X. Fu, J. Budzik, and K. J. Hammond. Mining navigation history for Recommendation. In Proc. 2000 international conf. intelligent user interfaces, pages 106-112, New Orleans, LA, January. ACM (2000)

[9] P.McJones. Eachmovie collaborative filtering data set. http://www.rearch. digital.com/SRC /eachmovie. DEC Systems Research Center (1997)

[10] 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 (1998)