추천을 위한 신경망 기반 협력적 여과
(Collaborative Filtering for Recommendation based on Neural Network)

김 은 주 † 류 정 우 †† 김 명 원 †††
(Eun Ju Kim) (Joung Woo Ryu) (Myung Won Kim)

요 약 추천은 과다하게 제공되는 정보로부터 사용자 개인의 취향에 알맞은 정보만을 제공하는 서비스이다. 최근 이러한 서비스는 정보제공자와 인터넷 사용자들이 많은 관심을 가지고 있다. 또한, 서비스를 위해 가장 널리 사용되는 방법은 협력적 여과방법이다. 협력적 여과방법은 특정 사용자와 관련 있는 사용자들에 대한 목표 항목의 선호도를 이용하거나 목표 항목과 관련 있는 항목들에 대한 특정 사용자의 선호도를 이용하여 특정 사용자에게 목표 항목을 추천하는 방법이다. 본 논문에서는 신경망 기반 협력적 여과 방법을 제안한다. 제안한 방법은 신경망을 이용하여 사용자 혹은 항목들 간의 선호 상관관계를 학습시킴으로써 모델을 생성하고 생성된 모델을 사용하여 추천할 목표 항목의 선호도를 추정하는 방법이다. 특히, 본 논문에서는 희소성 문제를 해결하기 위해 다양한 정보를 융합하는 방법과 보다 성능을 향상시키기 위해 목표 항목과 관련 있는 항목들 또는 특정 사용자와 관련 있는 사용자들을 선택하는 것에 대해 제시한다. 마지막으로 EachMovie 데이타를 이용한 실험들을 통해 제안한 방법이 기존 방법들 보다 우수한 성능을 보이는 것을 확인할 수 있다.

키워드 : 추천, 신경망, 협력적 여과, 정보 융합, 최근접 이웃 방법

Abstract Recommendation is to offer information which fits user's interests and tastes to provide better services and to reduce information overload. It recently draws attention upon Internet users and information providers. The collaborative filtering is one of the widely used methods for recommendation. It recommends an item to a user based on the reference users' preferences for the target item or the target user's preferences for the reference items. In this paper, we propose a neural network based collaborative filtering method. Our method builds a model by learning correlation between users or items using a multi-layer perceptron. We also investigate integration of diverse information to solve the sparsity problem and selecting the reference users or items based on similarity to improve performance. We finally demonstrate that our method outperforms the existing methods through experiments using the EachMovie data.

Key words : Recommendation, Neural Network, Collaborative Filtering, Data Fusion, Nearest Neighbor Method

1. 서 론

추천(recommendation)이란 과다하게 제공되는 정보로부터 사용자 개인의 취향이나 관심에 맞는 정보를 선택적으로 제공하는 서비스를 말한다. 이러한 서비스는 최근 전자상거래, 인터넷 쇼핑몰, 웹마이닝(web mining)등에 응용되고 있으나, 정확도에 있어 한계를 보이고 있다. 추천을 위한 여과 방법으로는 협력적 여과(collaborative filtering), 내용기반 여과(content-based filtering), 인구 통계학적 여과(demographic filtering)와 같이 크게 세 가지로 분류된다. 협력적 여과는 특정 사용자와 관련 있는 다른 사용자들에 대한 추천할 목표 항목의 선호도를 바탕으로 특정 사용자의 선호도를 추정하는 방법이며, 내용기반 여과는 항목의 다양한 속성 정보를 이용하여 사용자가 선호하는 항목을 추정하는 방법이다. 마지막으로 인구 통계학적 여과는 사용자의 나이, 성별, 직업 등 인구 통계학적 정보를 바탕으로 항목의 선

· 본 연구는 숭실대학교 교내 연구비 지원으로 이루어졌습니다.
† 학생회원 : 숭실대학원 컴퓨터학과
blue7786@bineee.pe.kr
†† 학생회원 : 숭실대학원 컴퓨터학과
ryu0914@orgio.net
††† 종신회원 : 숭실대학원 컴퓨터학과 교수
mkim@comp.soongsil.ac.kr
논문접수 : 2003년 4월 3일
심사완료 : 2004년 1월 6일

호도를 추정하는 방법을 말한다. 내용기반 여과는 논문, 전공서적과 같이 개인적 취향이나 전문성이 강한 항목 같은 경우 높은 추천 성능을 보일 수 있으며 인구통계학적 여과는 화장품, 의류와 같이 지역, 연령, 성별과 같은 특정 정보와 관련성이 많은 항목일 경우 높은 추천 성능을 보일 수 있다. 그러나 이러한 방법들은 추정하기 위한 정보들이 변하지 않는 정적인 정보이기 때문에 추천의 유연성이 부족한 문제점을 가지고 있으며 일반적으로 그 정확도가 협력적 여과보다 낮다[1-4].

기존의 협력적 여과 방법 중에서 대표적인 방법으로는 최근접 이웃 방법(nearest neighbor method)과 연관규칙 방법(association rule method) 등이 있다[5-7]. 최근접 이웃 방법은 적용하기가 용이한 반면 추천의 정확도가 낮다. 왜냐하면 추정하기 위해 사용자들 간의 유사도를 계산하는 데 있어 항목간의 중요도 즉, 가중치를 고려하지 못하기 때문이다. 반면, 연관규칙 방법은 다른 항목에 대한 선호도 자료에 적용하여 항목 선호에 대한 연관규칙들을 생성하고 그 규칙들을 사용하여 항목의 선호도를 추정하는 방법이다. 따라서 항목들 간의 중요도가 연관규칙의 지지도나 신뢰도로 나타난다고 할 수 있으나 단순히 항목들 간의 연관관계, 즉 표면적인 연관관계에 의하여 선호도를 추정함으로써 항목들 간의 의미적인 공통성 또는 상위 개념에 의한 선호도가 고려되지 않음으로써 역시 성능이 떨어지는 문제점이 있다[8,9].

본 논문에서는 신경망 기반 협력적 여과(CFNN : Collaborative Filtering based on Neural Network) 방법을 제안한다. 또한 제안한 방법의 성능을 향상시키기 위해 다양한 정보를 융합한 방법과 유사도를 이용하여 관련 있는 사용자 혹은 항목을 선택하는 방법을 제안한다. 제안한 방법은 신경망을 이용하여 사용자 혹은 항목들 간의 선호 상관관계를 학습시킴으로써 모델을 생성하고 그 모델을 사용하여 선호도를 추정하는 방법이다. 이 방법은 기존의 추천 방법이 가지고 있는 문제점들을 다음과 같은 장점으로 해결할 수 있다. 첫째, 항목이나 사용자들 간의 가중치를 학습할 수 있으므로 보다 정확한 선호도 계산이 가능할 뿐 아니라 신경망 은닉노드의 개념 형성 기능으로 보다 효율적인 선호도 산출이 가능하다. 둘째, 연속수치형, 이진 논리형, 범주형 등의 자료 유형에 상관없이 데이타의 처리가 용이하다. 셋째, 다른 내용, 인구 통계학적 정보 등 다른 이질적인 정보를 융합하기 용이하다. 특히, 기존 방법에서는 내용기반 여과 방법, 인구 통계학적 여과 방법이 각각 다른 방법으로 수행되는 데 비하여 신경망을 이용한 제안한 방법에서는 항목에 대한 내용 정보나 사용자의 인구 통계학적 정보를 단순히 입력 노드에 추가하여 학습시킴으로써 용이하게 융합할 수 있다.

본 논문은 다음과 같은 순서로 구성된다. 2장에서는 기존의 협력적 여과 방법들에 대해 설명하고 3장에서는 제안한 CFNN 방법과 성능을 향상시키기 위한 방법들에 대해 설명한다. 4장에서는 제안한 CFNN 방법의 성능을 향상시키기 위한 여러 가지 방법들의 성능 비교와 기존의 협력적 여과 방법들과 실험 결과를 비교한다. 마지막 5장에서는 결론을 맺고 향후 연구를 제시한다.

2. 관련 연구

협력적 여과는 사용자들 간의 상관관계를 찾아 이를 이용하여 관심 있는 정보들을 찾아내고 정보의 선호도를 추정하는 방법이다. 협력적 여과에서 사용자 간의 상관관계를 고려하는 방법은 여과의 결과에 큰 영향을 미친다. 따라서 이러한 방법에 대한 연구가 활발히 진행되고 있으며 그 중 대표적인 방법은 다음과 같다.

2.1 최근접 이웃 방법

GroupLens[10,11]에서는 처음으로 최근접 이웃 방법(k-nearest neighbor method)을 이용하여 사용자 간의 상관관계를 계산하였다. 즉 특정 사용자와 가장 가까운 k개의 근접 이웃(k-nearest neighbor)을 선택하여 다수결 원칙 또는 근접 정도에 따른 가중치 평균으로 목표 항목에 대한 특정 사용자의 선호도를 예측 계산하는 메모리 기반(memory-based) 협력적 여과 방법을 사용하였다. 메모리 기반이란 예측을 위해 전체 데이타들을 사용하는 것을 의미한다. 따라서 최근접 이웃 방법은 데이타가 증가할수록 수행속도 저하와 메모리가 증가되는 확장성(scalability) 문제를 가지고 있다[2].

$$c_{a,u}=\frac{\sum_{i=1}^{m}(r_{a,i}-\overline{r_a})(r_{u,i}-\overline{r_u})}{\sqrt{\sum_{i=1}^{m}(r_{a,i}-\overline{r_a})^2\sum_{i=1}^{m}(r_{u,i}-\overline{r_u})^2}} \quad (1)$$

$$p_{a,j}=\overline{r_a}+\frac{\sum_{u=1}^{k}[c_{a,u}(r_{u,j}-\overline{r_u})]}{\sum_{u=1}^{k}c_{a,u}} \quad (2)$$

최근접 이웃 방법은 사용자들 간의 상관관계를 피어슨 상관계수(Pearson correlation coefficient)로 계산한다. 피어슨 상관계수는 두 사용자 간의 선형 상관관계의 정도를 -1에서 1의 값으로 나타내며, 1에 가까워질수록 두 사용자 간의 양의 상관관계가 존재한다. 양의 상관관계란 사용자(a)가 선호하는 항목을 사용자(u)도 선호한다는 것을 의미한다. 반대로, -1에 가까워질수록 사용자(a)가 선호하는 항목을 사용자 (u)는 선호하지 않는다는 음의 상관관계가 있다는 것을 의미한다. 또한 상관계수가 0에 가까우면 선형 상관관계가 적다는 것을 나타낸다. 식 1에서 $c_{a,u}$는 사용자 a와 u의 피어슨 상관계수를 의미하며, $r_{a,i}$는 항목(i)에 대한 사용자(a)의

선호도를 나타내고, m은 사용자 a와 u가 동시에 선호도를 표시한 항목의 개수, $\overline{r_a}$, σ_a는 각각 사용자(a)가 선호도를 표시한 모든 항목들에 대한 선호도의 평균과 표준편차를 나타내고 있다. 식 2는 피어슨 상관계수를 가중치로 하여 사용자(a)의 추천할 목표항목(j)에 대한 선호도($p_{a,j}$)를 예측한다. k는 사용자(a)와 유사한 사용자 수이다.

이와 같이 최근접 이웃 방법을 이용한 협력적 여과에서는 사용자들 간의 상관관계를 계산하는 데 있어 항목의 중요도가 결여되어 있다는 것을 알 수 있다. 또한 적용이 쉬운 장점이 있으나 항목의 종류가 많은 데이타일 경우 희소성(sparsity) 문제가 발생할 수 있다. 희소성 문제란 사용자가 선호도를 표시한 항목의 개수가 적을 경우 사용자 간의 상관관계가 왜곡되는 문제를 말한다. 또한 항목의 종류뿐만 아니라 사용자 수가 많을 경우 수행속도가 느려지는 확장성(scalability) 문제도 고려해야 한다[5,8,9].

2.2 연관규칙 방법

데이타마이닝 기법 중 연관규칙 기법은 항목들의 빈도수와 동시 발생 확률을 이용하여 항목간의 관계를 찾아 연관성을 규칙으로 표현하는 기법이다. 연관규칙(규칙 R : IF A THEN B)에 대하여 지지도(*Support(R)*)는 전체 트랜잭션 중 A와 B항목이 동시에 일어난 트랜잭션의 비율을 의미한다. 연관규칙의 강도를 나타내는 신뢰도(*Confidence(R)*)는 조건 A를 만족하는 트랜잭션 중 B의 결론을 내릴 수 있는 트랜잭션의 비율을 의미한다.

$$Support(R) = \frac{A와\ B를\ 포함한\ 트랜잭션\ 수}{전체\ 트랜잭션\ 수} \quad (3)$$

$$Confidence(R) = \frac{A와\ B를\ 포함한\ 트랜잭션\ 수}{A를\ 포함한\ 트랜잭션\ 수} \quad (4)$$

빈발항목집합은 일정수준 이상의 지지도를 가지는 모든 항목들의 집합을 의미하며 빈발항목집합이 추출되면 이를 이용하여 연관규칙을 생성한다. 연관규칙 알고리즘에는 기본적으로 Apriori알고리즘이 사용한다. 이러한 연관규칙이 최근 추천시스템에서 많이 적용되고 있으나 [12]에서는 적합하지 않다고 서술하고 있다. 그 이유는 기존의 알고리즘들이 최소 신뢰도와 최소 지지도의 결정에 따라 생성되는 규칙의 개수가 다르기 때문에 사용자가 원하는 규칙의 개수를 제공하지 못하기 때문이다. 따라서 [12]에서는 최소 신뢰도에 대한 사용자가 원하는 규칙의 개수가 생성될 수 있게 최소 지지도를 자동으로 조정할 수 있고 사전에 선택된 한 개의 항목만이 규칙의 결론부에 나타나도록 알고리즘을 확장하였고, [7]에서는 [12]에서 제안한 알고리즘을 이용한 협력적 추천 방법을 제안하였다. 추천에 사용된 규칙의 형태는 아래와 같다. 규칙1은 사용자 연관규칙을 나타내고 있으며, *"추천 대상 항목에 대해 U_1이 선호하고 U_2가 선호하면 U_t도 선호한다."*는 것을 의미한다. 또한 규칙2는 항목에 대한 연관규칙을 나타낸다.

규칙1 : *IF* [U_1:선호] *AND* [U_2:선호]
THEN[U_t:선호] (지지도:70%, 신뢰도:85%)]

규칙2 : *IF* [I_1:선호] *AND* [I_2:선호]
THEN[I_t:선호] (지지도:60%, 신뢰도:90%)]

연관규칙의 경우 항목들이나 사용자들 간의 가중치가 연관규칙의 지지도나 신뢰도 등으로 나타난다고 할 수 있다. 그러나 단순히 표면적인 연관관계에 의하여 선호도를 결정함으로써 항목들 간의 어떤 내용적인 공통성 또는 어떤 상위 개념에 의한 선호도가 고려되지 않기 때문에 연관규칙 방법 역시 정확도가 떨어지는 문제점이 발생한다.

2.3 신경망 방법

[13]에서는 신경망 추천 방법을 제안한다. [13]에서 제안된 방법은 학습 알고리즘에서 결측값(missing value)을 처리하기 위해 그림 1과 같이 사용자 선호도를 학습데이타로 변환시킨다. 그림 1은 항목 I_5에 대한 사용자 U_t의 선호도를 예측하기 위해 U_t의 선호도가 표시된 항목들만을 선택하여 변환시킨 학습데이타를 보여준다. 사용자 선호도는 4등급(1,2 : 불호, 3,4 : 선호)으로 표시하고, 학습데이타는 사용자가 항목을 선호한다면 "10"으로 선호하지 않는다면 "01"로 나타낸다. 특히, 항목에 대한 선호를 알 수 없는 결측값의 경우 "00"으로 나타낸다. 결측값이 표현된 학습데이타에 SVD(Singular Value Decomposition)[14]를 적용하여 차원을 축약한 후, 항목에 대한 사용자들의 정보를 신경망의 입력으로 받고 U_t의 선호 정보를 신경망의 출력으로 하여 학습하는 방법이다.

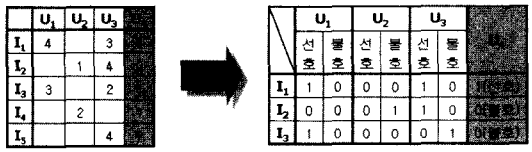

그림 1 선호도의 결측값 표현을 위한 표현

위에서 제시된 방법에서는 SVD을 사용함으로써 잡음을 제거하고 희소성 문제를 어느정도 해결하고 있으나 계산시간이 많이 걸리는 문제점을 가지고 있다.

2.4 베이지안 분류기 기반 방법

[15]에서는 유사도 기반 협력적 여과의 문제점을 해결하기 위하여 베이지안 분류기(bayesian classifier)을 이용한 협력적 여과를 제안한다. 이 방법에서는 신경망 방

법과 마찬가지로 결측값 처리를 위하여 그림 1과 같이 사용자 선호도를 학습 데이타로 변환시켜 사용하며, 사용자의 선호도를 선호와 불호로 나누어 계산한다.

베이지안 분류기 기반 방법에서는 단순 베이지안 분류기(simple bayesian classifier)의 표현을 위하여 변형된 데이타 모델(transformed data model)과 희소 데이타 모델(sparse data model) 2가지 모델을 정의한다. 변형된 데이타 모델이란 그림 1과 같은 변형된 데이타가 적용된 다변량 베르누이(multi-variate Bernoulli)모델과 동일하며, 모든 속성을 이용하여 모델을 생성한다. 희소 데이타 모델이란 아는 속성들만 이용하여 모델을 생성한다. EachMovie 데이타를 이용한 실험의 결과 변형된 데이타 모델보다 희소 데이타 모델의 성능이 더 좋게 나왔으며, 희소 데이타 모델은 유사도 기반 협력적 여과에 비하여 약 4%의 성능 향상을 보이고 있다.

베이지안 분류기 기반 방법은 유사도 기반 방법에 비하여 약간의 향상된 성능을 보이지만, 효과적인 속성 선택(feature selection)이 선행되어야 한다는 문제점이 있다.

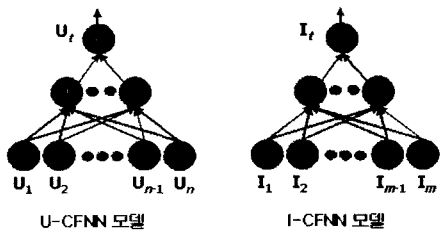

그림 2 신경망 기반 협력적 여과(CFNN)

3. 신경망 기반 협력적 여과

본 논문에서는 여러 응용분야에서 일반적으로 사용되는 기본 신경망 모델인 다층 퍼셉트론(MLP: Multi-Layer Perceptron)[16]을 사용한다. 본 장에서는 MLP를 기반으로 하는 협력적 여과 방법을 제안하고 성능향상을 위해 다양한 정보를 융합하는 방법과 관련있는 항목이나 사용자들을 선택하는 방법에 대해 설명한다.

3.1 사용자와 항목 신경망 기반 협력적 여과

제안한 방법은 추천 항목간의 선호 상관관계를 다층 퍼셉트론으로 학습시킨 모델을 사용하여 선호도를 추정하는 신경망 기반 협력적 여과(CFNN : Collaborative Filtering based on Neural Network) 방법이다. 이 방법은 그림 2와 같이 사용자 신경망 기반 협력적 여과(U-CFNN : User Collaborative Filtering based on Neural Network)모델과 항목 신경망 기반 협력적 여과(I-CFNN : Item Collaborative Filtering based on Neural Network)모델로 구분하여 생각할 수 있다. U-CFNN 모델은 사용자들 간의 상관관계를 바탕으로 특정 사용자에 대한 추천할 목표 항목의 선호도를 추정하고, I-CFNN 모델은 서로 다른 사용자들이 평가한 항목들 간의 상관관계를 바탕으로 추천할 목표 항목의 선호도를 추정한다.

즉, 그림 2의 왼쪽에서처럼 U-CFNN 모델은 특정 사용자인 목표 사용자(U_t)와 임의의 다른 사용자 <U_1, U_2, ..., U_{n-1}, U_n>들 간의 상관관계를 목표 사용자가 선호했던 항목들을 가지고 학습하여 생성된다. 추천할 목표 항목인 추천 항목이 주어지면 그 항목에 대한 다른 사용자들의 선호도를 입력으로 하여 추천 항목에 대한 목표 사용자의 선호도를 추정한다.

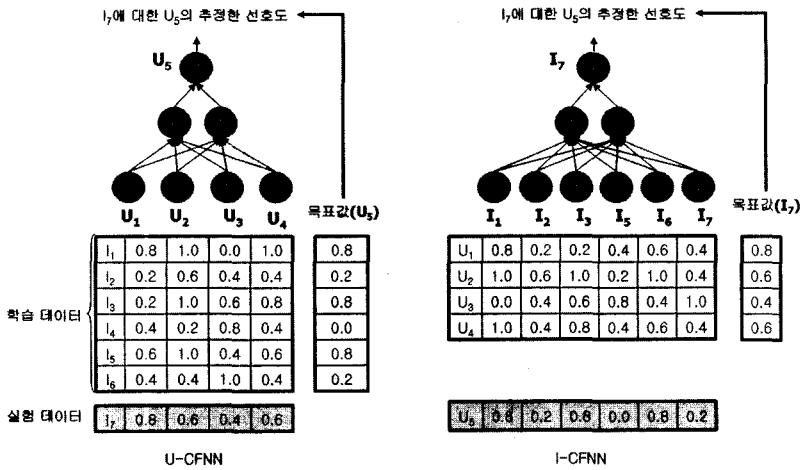

그림 3 CFNN의 학습의 예

표 1 5명의 사용자와 7개의 항목에 대한 선호도 정보 예제

	I_1	I_2	I_3	I_4	I_5	I_6	I_7
U_1	0.8	0.2	0.2	0.4	0.6	0.4	0.8
U_2	1.0	0.6	1.0	0.2	1.0	0.4	0.6
U_3	0.0	0.4	0.6	0.8	0.4	1.0	0.4
U_4	1.0	0.4	0.8	0.4	0.6	0.4	0.6
U_5	0.8	0.2	0.4	0.4	0.8	0.2	?

예를 들어, 표 1에서처럼 7개 항목과 5명 사용자의 선호도가 주어졌을 때, U-CFNN에 적용해 보면 그림 3의 왼쪽과 같이 목표 사용자(U_5)를 제외한 나머지 사용자들 <U_1, U_2, U_3, U_4>에 대한 선호도를 입력노드로 정의한다. 목표 사용자가 선호한 항목들 <I_1, I_2, I_3, I_4, I_5, I_6>에 대한 선호도를 학습데이타로 이용하여 U-CFNN 모델을 생성한다. 따라서 추천 항목(I_7)에 대한 목표 사용자의 선호도를 다른 4명의 사용자의 추천 항목에 대한 선호도를 이용하여 추정한다.

I-CFNN모델은 그림 3과 같이 U-CFNN모델과 구조는 같지만 학습데이타와 입력, 출력노드의 의미가 각각 다르다. 즉, U-CFNN 모델은 사용자간의 상관관계를 학습한 모델이고 I-CFNN은 항목간의 상관관계를 학습한 모델이다. 따라서 특정 사용자 혹은 목표 사용자에게 목표 항목을 추천하기 위해 U-CFNN 모델과 I-CFNN 모델 모두 사용할 수 있다. 특히, 효율적으로 사용하기 위해 사용자 수가 항목 수에 비하여 많다면 항목 모델인 I-CFNN을 이용하여 추천할 수 있으며, 반대로 항목의 수가 많다면 사용자 모델인 U-CFNN을 이용하여 추천할 수 있다. 또한 보다 정확한 추천을 위해서는 [17]과 같이 두 모델의 결과를 동시에 고려하여 사용할 수도 있다.

3.2 다양한 정보를 융합한 CFNN

협력적 여과 방법은 선호도가 공통으로 있는 항목 수가 적을 경우 추천 항목에 대한 선호도 추정의 정확성이 떨어진다. 즉, 어떠한 선호도 정보가 없는 초기 사용자와 같은 경우가 좋은 예시라고 말할 수 있다. 이와 같은 문제를 희소성(sparsity)문제라고 정의한다. 이러한 문제를 [12]에서는 SVD를 이용한 전처리 방법으로 해결하고 있으며, [1],[18]에서는 협력적 여과 방법과 내용기반 여과 방법을 융합하여 해결하고 있다. 특히 [18]에서는 베이시안 문서 분류기(Bayesian text classifier)를 이용하여 내용기반 예측기(content-based predictor)를 제안하고 있으며 이를 통해 선호도를 예측하고 예측된 선호도 정보를 이용하여 협력적 여과를 적용한 다음 추정하는 방법을 제안하고 있다. [1],[18] 모두 융합방법이 단일 방법 보다 성능이 향상되었으며 희소성 문제를 해결하고 있음을 보여준다.

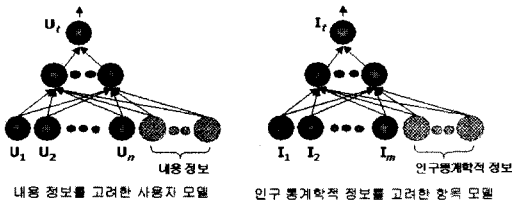

그림 4 다양한 정보를 융합한 CFNN

따라서 본 논문에서는 CFNN의 희소성 문제를 해결하기 위해 그림 4와 같이 다양한 정보를 융합한 CFNN을 설계한다. 목표 사용자가 선호했던 항목들을 학습데이타로 정의하고 있는 U-CFNN 모델에는 항목에 대한 내용정보를 융합하여 고려한다. 내용정보란 항목의 특성을 말한다. 예를 들면, 영화에 대한 내용정보를 영화장르, 감독, 주연배우 등으로 표현할 수 있다. 또한 추천 항목에 대한 선호도 정보를 가지고 있는 사용자들을 학습데이타로 정의하고 있는 I-CFNN 모델에는 사용자에 대한 인구 통계학적 정보를 융합한다. 인구 통계학적 정보란 사용자의 특성을 말하며 거주지, 성별, 직업, 취미 등으로 표현될 수 있다.

내용 정보를 융합하기 위하여 사용자 모델에 내용 정보의 수만큼 신경망 입력 노드를 추가하고 인구통계학적 정보를 융합하기 위하여 항목 모델에 인구통계학적 정보의 수만큼 신경망 입력 노드를 추가하여 학습시킨다. 특히, 기호(nominal)값을 처리하기 위해서는 기호값에 대응하는 입력 노드를 추가하여 0 또는 1의 값을 입력함으로써 해당 기호값을 표현한다. 예를 들어 사용자 모델에 내용 정보로 10개의 장르 정보를 융합한다면 사용자 모델에 입력 노드 10개를 추가하고 해당하는 장르를 1, 나머지 노드들의 값을 0으로 처리하여 모델을 생성한다.

이와 같이 기존의 방법들은 이질적인 정보를 융합하기 위해 전처리를 수행하거나 또는 다른 융합 방법을 고려해야하는 어려움이 있는 반면 신경망은 단지 입력 노드를 추가함으로써 쉽게 융합할 수 있는 장점을 가지고 있다.

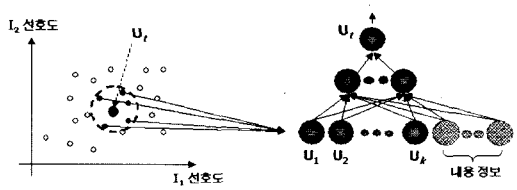

그림 5 유사한 k명을 고려한 CFNNS

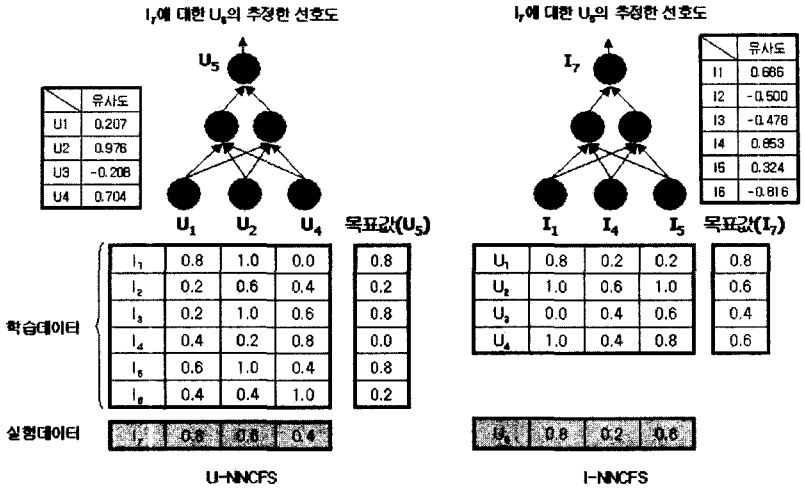

그림 6 CFNNS의 예

3.3 유사도를 이용한 신경망 기반 협력적 여과

앞에서 설명한 것과 같이 CFNN 방법은 다양한 정보를 융합함으로써 희소성문제를 해결하고 있지만 사용자 혹은 항목 수가 많아지면 입력노드가 증가되어 모델이 커짐으로 효율성이 떨어지는 문제를 가진다. 따라서 효율성을 높이기 위해 그림 5와 같이 목적사용자(U_t)와 연관있는 사용자를 선택하여 생성한 CFNN 모델을 고려할 수 있다. 앞으로 이러한 모델을 CFNNS(Collaborative Filtering based on Neural Network using Similarity)라고 표시한다. CFNNS는 유사도를 이용하여 목표 사용자와 연관있는 사용자 k명을 선택한다. 여기서 유사도는 피어슨 상관계수를 이용하여 계산되며 1에 가까운 k명을 말한다.

표 1의 예제를 CFNNS에 적용해 보면 그림 6과 같다. U-CFNNS은 U-CFNN와는 다르게 목적사용자(U_5)와 다른 사용자<U_1, U_2, U_3, U_4>간의 연관있는 항목들의 선호도를 이용하여 계산한다. 만약 k=3이라면 목적사용자와 가장 유사성이 높은 3명의 사용자 <U_1, U_2, U_4>을 선택하여 모델을 생성한다. 그림 7의 오른쪽에 있는 I-CFNNS도 추천 항목(I_7)과 다른 항목<I_1, I_2, I_3, I_4, I_5, I_7>간의 유사도를 계산하여 가장 연관있는 3개의 항목 <I_1, I_4, I_5>를 입력노드로 설정하고 모델을 생성한다. 유사도는 피어슨 상관계수 식 (1)을 이용하여 계산한다.

4. 실험 결과 및 분석

실험 데이타는 EachMovie 데이타[19]로써 72,916명의 사용자와 1,628개의 영화로 구성되어 있으며 각 고객이 본 영화에 대해서 평가한 선호도는 0.0, 0.2, 0.4, 0.6, 0.8, 1.0의 6단계 수치로 표현되어 있다. 본 실험에서는 [7]에서 제안하고 있는 연관규칙 방법과 성능을 비교하기 위해 같은 방법으로 최소 100편 이상 영화에 대한 선호도를 입력한 사용자 1,000명을 선택한다.

데이타에서 선호도가 입력되지 않은 경우를 처리하기 위해 표 2와 같이 데이타를 정량화하여 모델을 생성한다.

표 2 정량화 방법

선호도	0.0	0.2	0.4	0.6	0.8	1.0	선호도가 입력되지 않은 경우
정량화	-1.0	-0.6	-0.2	0.2	0.6	1.0	0.0

CFNN 모델을 생성하기 위해 실험에서 사용한 신경망 학습 파라미터는 실험분석을 통해 얻었다[16]. 학습속도와 최적해를 찾는데 밀접한 관계가 있는 학습률은 0.05로 설정한다. 신경망에서 학습은 MSE(Mean Square Error)가 0.0에 충분히 수렴할 때까지 학습을 시키는 것이 보통이지만 협력적 여과의 경우 어떤 특정 선호도 정보가 절대적인 것이 아니므로, 사용자들 간의 항목에 대한 선호도의 평균적인 경향을 학습하는 것이 중요하며 소수의 특정 데이타에 대하여 정확히 학습시킬 필요는 없다. 따라서 본 실험에서는 MSE가 0.04이하가 되면 학습을 종료시키고 은닉층 노드 개수를 5개로 설정하며 입력노드 개수는 100개로 한정한다.

본 실험에서 사용하는 성능 평가 척도는 기계학습에서 사용되는 Accuracy와 정보검색에서 사용되는 표준 척도인 Precision과 Recall 그리고 Precision과 Recall의 조화평균인 F-measure를 사용한다. 선호도가 표시된

전체 항목수를 T(Total Items), 그 중 사용자가 선호하는 항목의 개수가 TL(Total Like Items)이라고 할 때, 추천 시스템에 의해 사용자에게 추천되는 항목의 개수가 R(Recommended Items) 이고 그 중 사용자가 선호하는 항목의 개수를 RL(Recommended Like Items), 여기서 추천되지 않은 항목들 중에 사용자가 선호하지 않은 항목의 개수를 NRD(Not-Recommended Dislike Items)로 나타내면 성능 평가 척도는 다음과 같다.

$$Accuracy = \frac{RL + NRD}{T} \quad (5)$$

$$Precision = \frac{RL}{R} \quad (6)$$

$$Recall = \frac{RL}{TL} \quad (7)$$

$$F-measure = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (8)$$

식 5와 같이 Accuracy는 항목들을 선호, 불호로 얼마나 잘 분류되었는가를 나타내는 척도이다. Precision은 추천된 항목들 중에 사용자가 선호하는 항목들의 비율을 나타내며, Recall은 사용자가 선호하는 항목 수에 대한 추천된 항목 수의 비율을 나타낸다. 따라서 추천된 항목 수와 Recall은 비례적은 관계를 보이는 반면 Precision은 반비례적인 관계를 보인다. 그러므로 두 척도를 동시에 고려하기 위해 [13]에서는 식 8과 같이 두 척도에 대한 조화평균인 F-measure를 동시에 고려한다.

본 논문에서는 여러 가지 실험을 통해 제안하고 있는 CFNN 방법에 대한 타당성을 확인한다. 먼저 유사도를 이용한 CFNNS의 성능을 CFNN와 비교 분석하고, 다양한 정보를 융합한 성능을 평가한다. 마지막으로 기존 추천 방법과 비교한다.

4.1 CFNN와 CFNNS 비교 실험

본 실험에서는 선호도가 편향되지 않은 즉 선호 비율이 40~60%인 사용자 10명과 영화 10편을 선택하여 총 20개의 모델을 선택한다.

선택된 모델들은 4-fold cross validation으로 평가된다. 학습데이타의 구성은 사용자 모델의 경우 1000개의 속성으로 구성되어 있고, 항목모델은 1628개로 구성되어 있다. 모델을 생성할 때 목표항목을 제외한 모든 속성이 입력노드로 설정되어야 하지만 본 실험에서는 100개의 입력노드로 한정시킨다. 따라서 전체 사용자 속성 1000개 중에 100개 혹은 영화(항목) 1628개의 속성 중에서 100개를 선택하는데 있어 CFNN는 유사성을 고려하지 않고 앞에서 100개를 순차적으로 선택하여 모델을 생성하고 CFNNS는 피어슨 상관계수를 이용한 유사성을 고려하여 100개를 선택하고 모델을 생성한다.

표 3은 사용자, 항목 각각 10개에 대한 CFNN와 CFNNS의 실험 결과의 평균이다. 표에서 보는 바와 같이 항목간의 상관관계를 구하여 추천을 하는 것보다는 사용자간의 상관관계를 고려하여 추천하는 것이 보다 정확성이 높다. 또한 상관관계를 고려할 경우 임의로 사용자 혹은 항목을 선택하는 것보다 유사성을 고려하여 선택하는 것이 모든 척도에 있어 보다 높다는 것을 확인할 수 있다. 여기서 CFNN와 CFNNS의 성능의 차이가 사용자 모델에 비하여 항목 모델이 더 큰 이유는 항목 모델의 평균 유사도 차이가 사용자 모델의 평균 유사도 차이보다 크기 때문인 것으로 생각된다. 사용자 모델의 경우 CFNN일 때 평균 유사도가 0.27로 CFNNS일 때의 평균 유사도 0.54와 차이가 0.27인 것에 비하여 항목 모델의 경우는 CFNN의 평균 유사도가 0.21, CFNN-S의 평균 유사도가 0.53으로 0.32의 차이가 난다. 즉, 사용자 모델에 비하여 항목 모델의 유사도 증가분이 크기 때문에 관련성이 더 큰 항목을 선택하여 모델을 생성하게 되므로 성능 향상의 폭이 커진 것이라 설명할 수 있다.

표 3 CFNN와 CFNNS의 비교 실험 (단위:%)

	사용자 모델		항목 모델	
	U-CFNN	U-CFNNS	I-CFNN	I-CFNNS
Accuracy	82.1	83.6	77.7	81.5
Precision	81.7	82.6	77.9	82.0
Recall	84.1	85.6	75.8	79.4
F-measure	82.9	84.1	76.2	80.7

4.2 다양한 정보를 융합한 CFNN 실험

[1],[17]에서는 서로 다른 여과 방법을 사용하여 이질적인 데이타를 융합하였을 때 단일 방법만을 사용하였을 때 보다 향상된 성능을 보이며 특히 기존 여과 방법의 희소성 문제점을 해결하고 있다. 하지만 이들 융합 방법들은 원시적인 방법이거나 복잡한 방법을 적용함으로써 효율성이 낮다. 따라서 본 실험에서는 이질적인 데이타의 융합이 용이한 신경망의 장점을 이용하여 융합하였을 경우와 단일 데이타만을 사용하였을 경우를 비교하고 희소성 문제를 해결할 수 있는지 알아보기 위해 추정하기 위한 선호도 정보의 양을 조절하며 모델의 성능을 조사한다.

EachMovie 데이타에서 인구 통계학적 정보로는 성별, 나이, 지역, 직업이 있다. 지역과 직업에 대한 정보는 결측값이 너무 많아 사용하지 않았으며, 성별과 나이는 20대 남성이 약 40%이며, 특히 남성의 경우 약 83%를 차지하고 있어 판별력이 떨어짐으로 본 실험에서는 인구 통계학적 정보 융합에 대해서는 고려하지 않는다. 따라서 내용정보인 장르만을 융합하기 때문에 본 실험에서는 사용자간의 상관관계를 모델링한 U-CFNN만을

고려한다.

EachMovie 데이타에서 영화의 장르는 10개이다. 각각의 영화는 다수의 장르에 포함될 수 있으며 이진논리형으로 표현되어 있다. 제안된 모델에서는 융합할 장르의 정보에 가중치를 곱한 값을 입력값으로 설정하며 가중치는 사용자가 해당 영화에 대한 선호도로 정의한다. 이렇게 함으로써 사용자에 대한 장르의 선호도를 동시에 고려할 수 있다.

본 실험에서는 4.1과 같은 조건으로 사용자 10명에 대해 실험하였다.

표 4는 추정하기 위한 선호도 정보의 양 즉, 유사한 사용자 수를 조절하며 모델의 성능을 조사한 실험 결과이다. 유사한 선호도를 갖는 다른 사용자 수를 100명으로 하였을 때 장르를 고려하는 경우가 약간 성능이 향상된 결과를 보이고 있으나, 사용자 수가 줄어들수록 성능의 차이가 크게 나타남을 알 수 있다. 특히, 10명일 때, 장르를 고려한 경우 82.8%의 성능을 보이는 반면 고려하지 않은 경우 77.3%의 성능을 나타내며, 약 5% 성능 차이가 있음을 확인할 수 있다. 보통 입력 사용자 수가 증가함에 따라 성능이 향상되나 본 실험에서는 입력 사용자수가 50일 경우와 100일 경우가 비슷한 성능을 보였다. 이는 입력 사용자수에 상관없이 은닉층의 개수를 5개로 고정하여 실험하였기 때문에 입력 노드의 개수가 증가되어도 증가된 정보를 반영하지 못하기 때문이다.

따라서 다양한 정보를 융합하는 경우에는 유사한 선호도를 갖는 다른 사용자가 많은 경우에는 성능에 있어 크게 차이가 나지 않지만, 유사한 선호도를 갖는 다른 사용자가 적은 경우 즉, 초기 사용자와 같은 경우에 성능 차이가 큰 것을 알 수 있다. 그러므로 제안된 방법은 초기 사용자의 경우 다양한 정보를 융합함으로써 어느 정도 효율적으로 선호도를 추정할 수 있다.

표 4 입력 사용자 수별 U-CFNNS와 장르를 고려한 U-CFNNS 비교 (단위 : %)

	입력 사용자 수					
	10		50		100	
	사용자	사용자/장르	사용자	사용자/장르	사용자	사용자/장르
Accuracy	77.3	82.8	80.5	84.2	83.6	84.1
Precision	75.2	81.4	80.3	83.4	82.6	84.0
Recall	85.9	87.8	82.8	87.0	85.6	85.2
F-Measure	79.6	83.4	81.4	84.4	84.1	83.7

4.3 기존 추천 방법과의 성능 비교

기존 추천 방법과 비교하기 위하여, [7]과 같은 조건인 최소 100편 이상의 영화에 대한 선호도를 입력한 사용자 1,000명을 학습데이타로 사용하고, 70,000이상의 사용자 ID에서 최소 100편 이상 영화에 대한 선호도를 입력한 사용자 100명을 무작위로 선택하고 이를 검증데이타로 사용한다. 사용자와 항목 각각 30개씩 모델을 생성한다. 특히 사용자 경우 목적사용자 30명을 70,000이상의 ID를 갖는 사용자들 중 임의로 선택한다. 이때 20개 모델은 학습데이타가 편향되어 있지 않은 것을 선택하고 나머지 10개 모델은 선호 또는 불호로 학습데이타가 편향되어 있는 것을 선택한다.

표 5와 표 6은 생성된 각각 30개 모델에 대해 기존 추천 방법인 최근접 이웃 방법, 연관규칙 방법[6]과 제안하고 있는 CFNN와 CFNNS을 비교한 결과이다. 사용자 모델에서는 최근접 이웃 방법보다 본 논문에서 제안한 신경망 기반 협력적 여과 방법이 약 20% 향상되었고 항목 모델에서는 약 11% 향상되었다. 또한 신경망을 사용한 협력적 여과 방법 중 유사성을 고려한 CFNNS가 유사성을 고려하지 않은 CFNN보다 모든 척도와 모든 경우에 있어 성능이 향상된 것을 확인할 수 있다.

표 5 사용자 모델과 기존 추천 방법과의 비교(단위:%)

	최근접 이웃 방법	연관 규칙	U-CFNN		U-CFNNS	
			사용자	사용자/장르	사용자	사용자/장르
Accuracy	67.8	72.0	81.6	81.4	87.2	88.1
Precision	60.3	75.1	77.4	78.0	86.6	88.5
Recall	55.7	58.4	69.6	65.7	82.8	88.3
F-measure	57.9	65.7	73.3	71.3	83.3	85.7

표 6 항목 모델과 기존 추천 방법과의 비교 (단위 : %)

	최근접 이웃 방법	연관규칙	I-CFNN	I-CFNNS
Accuracy	68.0	61.1	76.5	79.1
Precision	71.1	75.4	76.1	74.4
Recall	60.6	22.6	72.1	76.6
F-measure	63.5	34.8	72.7	74.6

5. 결론 및 향후 연구

본 논문에서는 신경망 기반 협력적 여과(CFNN) 방법을 제안하고, 제안한 방법의 성능을 향상시킬 수 있는 방법들을 제안한다. 성능을 향상시키기 위한 방법으로는 다양한 정보를 추가하는 방법과 유사성을 고려하여 특정 사용자 혹은 목표 항목과 관련 있는 사용자와 항목들을 선택하는 방법이다. 신경망 기반 협력적 여과 방법은 학습을 통해 데이타 간의 복잡한 관계를 고려할 수 있고 다양한 정보를 효율적으로 융합할 수 있는 신경망

의 장점을 활용하였다. 기존 협력적 여과 방법과의 비교 실험을 통해 제안한 방법의 성능이 우수함을 확인할 수 있었다. 제안한 방법은 결과에 대한 이유를 설명할 수 없는 문제점을 가지고 있지만 이는 추천에서 중요하지 않다.

향후에는 보다 추천의 성능을 향상시키기 위해 사용자 행동 패턴과 같은 다양한 정보들을 융합하는 방법에 대해 연구할 것이다. 또한 지식을 효율적으로 표현할 수 있는 규칙기반 모델과 통합할 수 있는 방법에 대해 연구할 것이다. 특히, 신경망 기반 협력적 여과 모델에서 은닉층 노드들의 역할 분석에 관심을 두고 있다. 따라서 규칙기반 모델과 통합된 모델은 은닉층 노드들의 숨겨진 개념적 구조의 역할을 나타낼 수 있을 것으로 기대된다. 마지막으로 제안한 방법이 실생활에 적용될 수 있는지 더 많은 데이타들을 이용하여 확인할 것이다.

참 고 문 헌

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

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

[3] Schafer, J. B,. Konstan J. A., and Ried, J., "E-commerce Recommendation Applications," J. Data Mining and Knowledge Discovery, 2001.

[4] Cheung, K. W., Kwok, J. T., Law M. H. and Tsui, K. C., "Mining customer product ratings for personalized marketing," Decision Support Systems, Volume 35, Issue 2, pp. 231-243, 2003.

[5] 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.

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

[7] Lin, W., Ruiz, C., and Alverez, S. "A 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), 2000.

[8] Herlocker, J. L., Konstan, J. A., Borchers, A., and Riedl, J., "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.

[9] Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., and Sartin, M., "Combining Content-Based and Collavorative Filters in an Online Newspaper," In Proceedings of ACM SIGIR'99 Workshop in Recommender Systems : Algorithms and Evaluation, Univ. of California, Derkely, Aug, 1999.

[10] Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., and Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J., "GroupLens: An Open Architecture for Collaborative filtering of Netnews," In proceddings of CSCW '94, Chapel Hill, NC, 1994.

[11] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J., "GroupLens: Applying Collaborative Filtering to Usenet News", Communications of the ACM, 40(3), pp. 77-87, 1997.

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

[13] Billsus, D., and Pazzani, M. J., "Learning collaborative information filters," In Proceedings of the Fifteenth International Conference on Machine Learing, pages 46-53, 1998.

[14] Press, W. H., Teukolsky, S. A., Vetterling W. T., Flannery, B. P., "Numerical Recipes in C++," 2nd edition, Cambridge University Press, 2002.

[15] Miyahara, K., Pazzani, M, J,. "Collaoborative Filtering with the Simple Bayesian Classifier," Pacific Rim International Conference on Artificial Intelligence, Springer, pp 679-689, 2000.

[16] Haykin, S., "Neural Networks : A Comprehensive Foundation," 2nd edition, Prentice Hall, 1999.

[17] 도영아, 김종수, 류정우, 김명원, "협력적 추천을 위한 사용자와 항목 모델의 효율적인 통합 방법", 한국정보과학회, Vol. 30, No 5·6, pp. 540-549, 2003.

[18] Prem Melville, Raymond J. Mooney, and Ramadass Nagarajan, "Content-Boosted Collaborative Filtering," Proceeding of the SIGIR-2001 Workshop on Recommender Systems, New Orleans, LA, Sep., 2001.

[19] P. McJones. "Eachmovie Collaborative Filtering Data set," http://www.rearchdigital.com /SRC/eachmovie, DEC Systems Research Center, 1997.

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

김 은 주

2001년 2월 숭실대학교 자연과학대학 정보통계학과 졸업(학사). 2003년 2월 숭실대학교대학원 컴퓨터학과 졸업(석사). 2003년 3월~현재 숭실대학교대학원 컴퓨터학과 박사과정. 관심분야는 데이타마이닝, 웹마이닝, 추천 시스템, 신경망, 에이전트, 시맨틱 웹

류 정 우

1998년 2월 숭실대학교 정보과학대학 인공지능학과 졸업(학사). 2000년 2월 숭실대학교대학원 컴퓨터학과 졸업(석사). 2000년 2월~현재 숭실대학교대학원 컴퓨터학과 박사과정. 관심분야는 신경망, 유전자알고리즘, 퍼지이론, 데이터마이닝, 에이전트

김 명 원

1972년 서울대학교 응용수학과 졸업. 1981년 University of Massachusetts (Amherst) Computer Science 석사 학위 취득. 1986년 University of Texas (Austin) Computer Science 박사 학위 취득. 1975년~1978년 한국과학기술연구소 연구원. 1982년~1985년 Institute for Computing Science & Computer Application(Univ. of Texas) 연구원. 1985년~1987년 AT&T Bell Labs.(Naperville) 연구원. 1987년~1994년 한국전자통신연구소 책임연구원. 1992년~1993년 한국신경회로망 연구회 회장. 1993년~1995년 IEEE Neural Network Council 한국지부장. 1993년~1995년 정보과학회 뉴로컴퓨팅연구회 위원장. 1994년~현재 숭실대학교 컴퓨터학부 교수. 1998년~2000년 한국인지과학회 부회장. 2000년~2001년 미국 IBM T.J Watson 연구소 방문과학자 2001년~2002년 한국뇌학회 회장. 관심분야는 신경회로망, 퍼지시스템, 진화알고리즘, 패턴인식, 자동추론, 기계학습, 데이타마이닝, creativity engineering 등