한국지능정보시스템학회논문지 제9권 제3호 2003년 12월(pp.117~135)

자기 조직화 신경망(SOM)을 이용한 협력적 여과 기법의 웹 개인화 시스템에 대한 연구

강부식
목원대학교 경영정보학과
(bookang@mokwon.ac.kr)

개인화된 정보를 제공하기 위한 협력 여과 기법에 대한 많은 연구가 이루어지고 있는 데, 유사 사용자들을 찾는 과정에서 상관계수와 같은 유사성 척도를 이용하여 모든 사용자와의 유사성을 계산하는 과정을 거친다. 이 때 사용자 수가 많아지게 되면, 계산의 복잡도가 지수적으로 증가하게 되는 규모의 문제가 발생한다.

본 연구는 협력 여과 기법에서 주로 사용하는 유사성 척도가 사용자 집단이 커짐에 따라 계산의 복잡도가 지수적으로 증가하는 문제를 해결하기 위한 방안을 제시하는 것이 주 목적이다. 규모의 문제를 해결하기 위해 클러스터링 모델 기반 접근 방식을 사용하고 아이템의 선호도 계산을 위해 RFM(Recency, Frequency, Momentary) 기준의 사용을 제안한다. 먼저 SOM을 이용하여 전체 사용자를 사용자 집단으로 클러스터링하고 사용자 집단별로 RFM 기준에 의해 아이템의 점수를 계산하여 선호도가 높은 순으로 정렬하여 저장한다. 사용자가 로그인하면 학습된 SOM을 이용하여 대상 사용자 집단을 선정하고 미리 저장된 추천 아이템을 추천한다. 추천결과에 대해 사용자가 평가하면 그 결과를 이용하여 현 시스템의 개정 여부를 결정한다. 제안한 방안에 대해 MovieLens 데이터 셋에 적용하여 실험한 결과 기존의 협력적 여과 기법에 비해 추천 성능이 비교적 우수하면서도 추천 시스템 운용시의 계산 복잡도를 일정하게 유지시킬 수 있음을 보였다.

논문접수일 : 2003년 7월 게재확정일 : 2003년 12월 교신저자 : 강부식

1. 서론

고객과의 지속적인 관계를 통해 고객의 평생 가치를 극대화하기 위한 관계마케팅은 기업의 경쟁이 심화될수록 그 중요성이 더해 가고 있다. 인터넷의 급속한 발전에 따라 인터넷을 이용한 고객과의 관계마케팅이 점차 증가하고 있는 데, 이때 중요한 역할을 담당하는 기술중의 하나가 웹 개인화 기술이다. 개인화는 목표 고객을 선정하고, 고객의 친밀도를 증가시키며, 브랜드에 대한 고객의 충성도를 강화시켜서 궁극적으로는 매출로 이어지게 한다. 웹 개인화는 서비스, 정보, 광고, 상품 등에 대해 각 고객에게 필요한 맞춤화된 정보를 제공한다. 인터넷 웹사이트에서 가용한 정보의 양이 많아지면 많아질수록, 정보의 홍수속에서 서비스와 상품에 대한 개인화된 정보의 제공은 더욱 필요하다.

웹 개인화는 많은 방법에 의해 구현될 수 있는

* 이 논문은 2002년도 한국학술진흥재단의 지원에 의하여 연구되었음(KRF-2002-003-B00057).

데, 웹 개인화의 주요 기술중의 하나가 정보 여과 기법이다. 정보 여과 기법은 크게 내용 기반 여과 기법(content-based filtering)과 협력적 여과 기법(collaborative filtering)으로 나눌 수 있다(Lee et al., 2001; Kohrs and Merialdo, 2001). 내용 기반 접근법은 과거에 대상 고객이 선호했던 아이템과 가장 유사한 아이템을 찾아 개인화된 정보를 제공한다. 즉 아이템과 아이템 사이의 관계를 토대로 정보를 제공한다. 협력적 여과 기법은 사용자간의 연관성을 기반으로 개인화된 정보를 제공하는 데, 대상 고객과 유사한 성격을 가진 다른 고객들이 선호한 것으로 밝혀진 아이템을 추천하는 방식이다.

협력적 여과 기법 기법은 오늘날 추천 시스템 분야의 가장 핵심적인 기법으로 위치하고 있으며 인터넷에서 괄목할 만한 성공을 보이고 있다(Konstan et al., 1997; Shardanand and Maes, 1995; Hill et al., 1995; Goldberg et al., 1992).

협력적 여과 기법에서 가장 일반화된 알고리즘은 이웃-기반 기법이다. 이웃 기반 기법에서 대상 고객과 유사성이 높은 고객군이 선택되고 대상 고객에 대한 추천을 위해 유사 고객군의 아이템들에 대한 선호도를 반영하여 추천 아이템을 결정하게 된다. 이웃 기반 기법은 3 단계로 구분할 수 있다(Herlocker et al., 1999).

1) 대상 고객과 모든 고객들을 유사성 척도를 이용하여 측정한다.
2) 대상 고객을 위해 예측에 반영할 유사 고객군을 선정한다.
3) 유사 고객군의 선호 아이템을 중심으로 대상 고객에게 아이템을 추천한다.

협력적 여과 기법을 정리해 보면 크게 두가지 기능을 수행한다. 먼저 대상 고객에 대한 유사 고객군을 찾아내고, 다음에 유사 고객군의 선호도가 높은 아이템을 추천하는 것이다. 비슷한 성향을 갖는 다른 고객들을 찾음에 있어 일반적으로 대상 고객과 전체 고객간의 유사성을 매번 구하여 유사 고객군을 선정하고 있는 데, 고객 집단이 커지면 연산처리가 기하급수적으로 늘어나는 규모의 문제(scalability problem)가 발생하게 된다. 또한 대부분의 추천시스템이 추천 그 자체에만 관심이 있고 추천 결과의 반영에 대해서는 언급하고 있지 않다.

이 연구에서는 협력적 여과 기법을 적용할 경우 발생하는 규모의 문제를 해결하기 위해 신경망의 자기조직화맵(SOM: Self-Organizing Map) 알고리즘(Kohonen, 1988)을 이용하는 방안에 대해 제안하고자 한다. SOM을 이용하여 고객을 세분화된 고객군 그룹으로 클러스터링하고, 대상 고객이 속하는 고객군의 선호도가 높은 상품을 추천하는 방안에 대해 제안한다. 이렇게 함으로써 시스템 운용시점에서의 계산의 복잡도를 일정하게 유지할 수 있게 된다. 그리고 추천한 상품에 대한 고객의 만족도 평점을 반영하여 추천 시스템을 보정하기 위한 시점을 제시하고자 한다. 이 연구의 구성은 다음과 같다. 2장에서는 웹 개인화 서비스를 지원하기 위해 협력적 여과 기법을 사용한 연구와 자기 조직화 신경망, RFM(Recency, Frequency, Monetary)기법에 대해 살펴본다. 3장에서는 이 연구에서 제안하는 자기 조직화 신경망을 이용한 협력적 여과 기법의 프레임웍을 제시한다. 4장에서는 제안한 방안을 MovieLens[1] 데이터 셋에 적용한 실험결과를 나타내고 분석한다. 5장은 연구의 결론을 제시한다.

1) http://movielens.umn.edu/

2. 관련 연구

2.1 웹 개인화를 위한 협력적 여과 기법

웹 개인화를 위한 내용 기반 여과 기법과 협력적 여과 기법에 대해 살펴보면 다음과 같은 특징이 있다(안현철, 2002). 내용기반 기법의 장점은 상품 자체를 모델링하는 기법이기 때문에 직접적이고 단순하다. 또한 전반적인 고객군이 이질적인 평가를 보이는 상품군의 추천에 있어 좋은 예측력을 보인다. 내용 기반 여과 기법의 두가지 타고난 한계점으로는 먼저, 각 아이템에 대한 특성을 추출하고 이를 기반으로 추천 대상을 정하여야 하는데 이것은 사실상 효과적으로 이루어지기 어렵다는 점이다. Herlocker등(1999)은 멀티미디어 컨텐츠로 구성된 웹사이트의 내용기반 분석은 매우 어려움을 지적하고 있다. 두번째는 내용기반 기법은 고객이 이전에 좋게 평가한 상품과 유사한 상품군을 찾기 때문에, 추천 결과가 너무 과도하게 특정 부분에 치우치게 된다는 점(over specialization)을 지적한다. 협력적 여과 기법의 장점은 새로운 상품군에 대해서도 추천이 가능하고, 일반적으로 고객군이 동질한 평가를 보이는 상품군에 대해 상대적으로 높은 예측력을 보이고(Konstan et al., 1997), 데이터가 충분한 경우 다른 기법에 비해 상대적으로 높은 예측력을 보인다는 점이다. 주요 한계점으로는 고객의 구매이력과 아이템 선호도로만 모델링되므로, 고객의 구체적인 프로필 정보는 반영되지 못하고 있다는 점과 초기에 많은 정보를 필요로 하고, 전반적인 고객군이 서로 이질적인 평가를 보이는 상품군에 대해서는 예측력이 떨어지고, 고객군의 크기가 커질수록 많은 연산처리를 필요로 한다는 점을 들고 있다.

협력적 여과 기법은 오늘날 추천 시스템 분야의 가장 핵심적인 기법으로 위치하고 있으며, 협력적 여과 시스템은 Usenet 뉴스 기사 선정을 도와주는 GroupLens(Konstan et al., 1997), 영화 추천을 위한 MovieLens, Ringo 음악 추천(Shardanand and Maes, 1995), Bellcore 비디오 추천(Hill et al., 1995) 등 여러 영역에 적용되어져 왔다.

그러나 협력적 여과 기법의 한계점은 상존하고 있는데, 이를 해결하기 위해 Lee등(2001)은 협력적 여과 기법과 연관(association) 규칙 마이닝 기법을 결합한 웹 개인화 시스템의 프레임웍을 제안하였고, Cohen과 Fan(2000)은 자동화된 웹 스파이더(Web spider)를 활용하여 협력적 여과 기법을 효과적으로 할 수 있음을 보였다. Kohrs와 Merialdo(2001)는 사용자에게 웹 박물관에서 맞춤화된 웹 페이지를 제공할 수 있도록 협력적 여과 기법과 내용 기반 접근 방식을 결합하여 적용하는 방안에 대해 보였다. Lee등(2003)은 퍼지 연관 메모리(fuzzy association memory)를 적용하여 협력적 여과 기법을 하는 시스템에 대해 연구하였고, Resnick과 Varian(1997)은 여러 다양한 추천시스템을 제시하고 있다. 그러나 위의 연구들에서 유사도 측정을 위해 상관계수와 기타 유사한 척도를 사용함으로써 규모의 문제는 계속 발생하고 있다.

전통적인 협력적 여과 기법은 상관계수를 이용하는 것으로, 대표적인 예가 GroupLens (Konstan et al., 1997) 시스템이다. GropuLens는 이웃 기반 알고리즘을 사용하여 처음으로 자동화된 협력적 여과 기법을 적용하였다. 본래의 GroupLens 시스템은 피어슨의 상관계수를 사용하여 사용자들의 유사성을 측정하였고, 유사 사용자들의 가중화된 평점을 이용하여 다음 식(1)

같이 예측을 하였다.

$$p_{a,i} = \bar{r}_a + \frac{\sum_{u=1}^{n} (r_{u,i} - \bar{r}_u) \times w_{a,u}}{\sum_{u=1}^{n} w_{a,u}} \quad (1)$$

Pa,i는 대상 사용자 a를 위한 아이템 i의 예측값을 나타낸다. n은 이웃의 수를, Wa,u는 대상 사용자 a와 이웃 u사이의 유사성 가중치를 나타내는데 피어슨의 상관계수로 표현된다.

$$w_{a,u} = \frac{\sum_{i=1}^{m} (r_{a,i} - \bar{r}_a) \times (r_{u,i} - \bar{r}_u)}{\sigma_a \times \sigma_u} \quad (2)$$

상관계수 및 이와 유사한 척도를 사용하게 되면 고객집단이 커짐에 따라 유사성 측정을 위해 많은 연산이 요구되게 되고 결국 규모의 문제가 발생하게 된다.

규모의 문제를 해결하기 위한 한 가지 대안은 모델 기반 접근 방식을 취하는 것이다(안현철, 2002; Sarwar et al., 2001; Sarwar et al., 2000). 모델 기반 방식에서는 대상 고객에게 아이템을 추천하기 위해 전체 고객 정보를 탐색하는 전통적인 방식 대신, 기존의 데이터들을 토대로 한 번 추천을 위한 모델이 완성되면, 이후에는 미리 구축한 모델을 이용하여 간단한 연산만으로 추천 결과를 만들어 냄으로써 적은 컴퓨터의 연산요량만으로도 추천 아이템을 제시할 수 있다. 모델 기반 접근 방식의 사례로는 Bayesian network (Breese et al., 1998), 클러스터링 기법(Ungar and Foster, 1998)과 같은 기계학습 알고리즘이 사용된 연구가 소개되고 있다. Bayesian network 기법은 의사결정 나무기법을 이용하여 모델을 구축한다. 구축된 모델은 매우 작고, 빠르고, 이웃 기반 기법과 거의 유사한 성능을 보인다. Bayesian network 기법은 고객의 선호가 시간에 따라 아주 느리게 변하는 환경에서 효과적이나 고객의 선호가 빠르게 변화하는 환경에서는 적절치 않은 것으로 평가받고 있다(Sarwar et al., 2001). 클러스터링 기법은 전체 고객을 비슷한 성향을 보이는 고객군으로 나누고, 대상 고객이 속하는 고객군의 선호 아이템을 이용하여 추천하는 방식으로 추천 시스템의 연산 성능을 매우 향상시키나 이웃 기반 방식에 비해 추천 성능이 떨어지는 것으로 보고되었다(Sarwar et al., 2001; Sarwar et al., 2002). 협력적 여과 기법을 사용한 추천 시스템에 대한 많은 연구 결과(Sarwar et al., 2001; Sarwar et al., 2000; Herlocker et al., 1999; Konstan et al., 1997)를 발표하고 있는 미네소타 대학의 GroupLens 연구팀은 규모의 문제를 해결하기 위해 모델 기반 접근방식을 사용하는 방안을 제시하였다(Sarwar et al., 2000; Sarwar et al., 2001). Sarwar등(2000)과 Gupta와 Goldberg(1999)는 SVD (Singular Value Decomposition) 기법을 이용하여 계산의 공간을 사전에 줄임으로써, 온라인에서 매우 빠르게 추천할 수 있음을 보였다. 그러나 MovieLens 데이터셋의 실험에서 Sarwar등(2000)은 전통적인 협력적 여과 기법에 비해 SVD를 이용한 기법의 추천 성능이 항상 우수하지는 못함을 실험적으로 보여주고 있다. 다만 응용 분야에 따라 SVD 적용의 잠재력이 있을 수 있음을 주장하고 있다. Sarwar등(2001)의 연구에서는 사용자 간의 유사성을 측정하는 대신 아이템간의 유사성을 사전에 측정하여 저장해 두었다가 사용자에게 이 정보를 이용하여 아이템을 추천하는 방안을 제시하고 있으며, GroupLens 연구팀의 MovieLens 데이터셋에 대한 실험에서 규모의 문제를 해결하면서 추천 성능도 우수함을

실험적으로 보여주고 있다. 이 연구는 협력적 여과 기법을 해결하는 한 방안으로 내용 기반 여과 기법을 사용한 것으로 내용 기반 여과 기법이 가지는 한계를 갖게 된다.

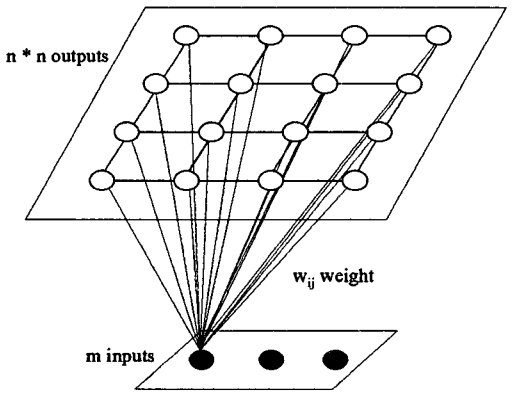
본 연구는 사용자간의 유사성을 이용한 협력적 여과 기법을 이용한 규모의 문제 해결에 초점을 두고 있다. 전체 사용자간의 유사성을 실시간으로 측정하는 전통적인 방식 대신 특성이 유사한 사용자 집단을 미리 군집화한 후에 이 정보를 이용하게 되면 실시간으로 계산하는 처리시간을 줄일 수 있다. 사용자 집단으로 사전에 군집화하기 위해서 클러스터링 기법을 사용한다. 클러스터링 기법은 마케팅에서 고객세분화를 위한 대표적인 기법으로 활용되고 있다. Sarwar등(2002)은 k-means 클러스터링 기법을 이용한 경우 규모의 문제는 해결하지만 추천 성능이 떨어짐으로 시스템의 빠른 처리 능력과 추천 성능간의 교환(trade-off) 의사결정이 필요하다고 지적하고 있다.

이 연구에서는 클러스터링 기법으로 SOM을 이용하여 추천 성능이 떨어지지 않으면서 협력적 여과 기법의 규모 문제를 해결할 수 있음을 보이고자 한다. SOM은 Kohonen(1988)이 제안한 신경망 알고리즘으로 클러스터링 문제에 효과적임을 여러 연구에서 밝히고 있는데, Smith와 Ng (2003)는 웹 페이지의 클러스터링을 위해 SOM을 사용하는 방안을 제시하였고, Changchien과 Lu(2001)는 온라인으로 추천하는 시스템에서 클러스터링 기법으로 SOM을 사용하는 방안을 제시하였다. 이 연구에서도 클러스터링 기법으로 SOM을 이용하여, 전체 고객을 성향이 유사한 고객군으로 분리하고 고객군의 선호도를 바탕으로 대상 고객에게 선호도가 높은 아이템을 추천하는 방안을 제시하고자 한다. 고객군내에서 각 아이템에 대한 선호도를 결정하는 방식은 마케팅 분야에서 많이 사용하고 있는 RFM (Recency, Frequency, Momentary)(Bult and Wansbeek, 1995; Ha and Park, 1998) 기준을 적용하고자 한다.

2.2 자기 조직화 신경망(SOM)과 RFM 기준

SOM신경망은 Kohonen(1988)이 제안한 무감독 학습기법으로, <그림 1>과 같이 구성된다. SOM은 입력층과 출력층으로 구성된 신경망으로, 입력층의 입력노드수는 입력벡터의 차원의 수 m으로, 출력층의 노드 수는 n*n의 격자모양으로 구성한다. 학습은 입력층과 출력층을 연결하는 연결가중치의 값을 조정하는 것이다.

학습이 끝나면 비슷한 성격의 입력벡터 값을 가지는 데이터들은 출력층의 특정노드에 맵핑이 된다. 따라서 비슷한 성격의 입력 데이터는 자연스럽게 동일한 클러스터링으로 구성될 수 있다. 또한 SOM의 특성중 하나는 SOM을 학습시에 학습 데이터간의 관계에 대한 사전지식이 없어도 SOM의 학습 규칙에 따라 연결가중치가 학습이 되기 때문에 복잡한 연관관계가 있는 데이터 셋



<그림 1> 자기 조직화 신경망의 구성 형태

에 대해 쉽게 학습이 가능한 점이다.

SOM의 학습 알고리즘은 내부에서 사용하는 유사성의 척도 및 이웃함수, 그리고 학습 파라미터의 결정 등에 따라 약간의 차이가 있으나 다음과 같이 이루어 진다(Bigus and Bigus, 2001).

먼저, 입력 데이터 셋이 입력층에 전달된다. 다음에, 입력벡터 x와 출력노드 j를 연결하고 있는 연결가중치 벡터 w_j와의 거리를 유클리디안 거리로 계산한다.

$$y_j = \|x - w_j\|^2 \tag{3}$$

가장 적은 y_j을 가진 출력노드 j를 승리 노드로 선정한다. 승리 노드와 승리 노드의 이웃 노드의 연결가중치를 다음 식을 따라 조정한다.

$$w_j(t+1) = w_j(t) + \beta(k) C_{ij}(k) y_j(t) \tag{4}$$

위 식(4)에서 β(k)는 학습 반복 k 단계에서의 학습율을, $C_{ij}(k)$는 학습 반복 k단계에서 노드 i와 j의 이웃 함수 값을, y_j는 t 시점에서 입력벡터 x와 가중치 벡터 w_j사이의 유클리디안 거리를 나타낸다. $C_{ij}(k)$는 Gaussian 함수로 멕시칸 모자 형태를 취하고 있는데, 다음과 같이 정의된다.

$$C_{ij}(k) = \exp\left(-\frac{\|i-j\|^2}{\sigma(k)^2}\right) \tag{5}$$

위 식(5)에서 i, j는 SOM의 n*n 2차원 출력노드 맵에서의 좌표, 즉 출력노드 i와 j의 좌표를 나타내고, k는 학습 단계를 나타낸다. 이웃 함수의 넓이는 $\sigma(k)^2$으로 학습 단계 k가 적을 때는 출력노드 전체에 해당할 만큼 넓게, k가 최대값에 이르면 출력노드 중 한 노드만 해당이 되도록 정의한다. 즉 학습이 진행될수록 점차 이웃함수의 범위를 좁히게 된다.

β(k)는 다음 식(6)와 같이 계산된다.

$$\beta(k) = \beta_{initial}\left(\frac{\beta_{final}}{\beta_{initial}}\right)^{\frac{k}{k_{max}}} \tag{6}$$

k_{max}는 수행해야 할 최대 학습 반복회수를 의미한다. 학습율 β(k)는 k가 증가함에 따라 지수적으로 감소한다. 일반적으로 $\beta_{initial}$=1.0을, β_{final}=0.05를 사용한다.

마케팅의 고객 분류를 위해 많이 사용하는 기준 중의 하나인 RFM(Recency, Frequency, Momentary)(Bult and Wansbeek, 1995; Ha and Park, 1998) 기준은 다음과 같은 3가지 기준을 사용하여 고객을 분류한다.

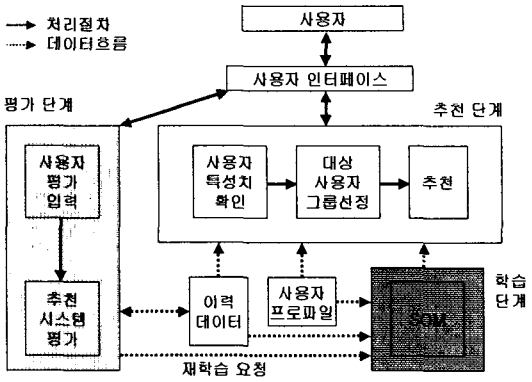
R(Recency): 아이템에 대한 이벤트의 발생 시기가 최근일수록 높은 점수를 부여하는 기준이다. 예를 들어 상품 판매 사이트에서 최근에 구매된 상품일수록 높은 점수를 부여한다.

F(Frequency): 주어진 기간 내에 아이템과 관련하여 발생한 이벤트의 빈도 수가 높을 수록 높은 점수를 부여한다. 상품 판매 사이트라면 주어진 기간에 사용자가 구매한 횟수가 많을수록 높은 점수를 부여한다.

M(Momentary): 주어진 기간 내에 아이템의 총 구매금액이 클수록 높은 점수를 부여한다. 상품 판매 사이트에서 일정한 기간 내에 사용자가 구매한 금액이 클수록 높은 점수를 부여하는 기준이다.

3. SOM을 이용한 개인화 방안

웹 개인화를 위한 협력적 여과 기법 시스템에 SOM을 활용하기 위한 연구는 크게 3가지 단계로 구성된다. 사용자 그룹을 세분화된 그룹으로 클러스터링하기 위해 SOM을 학습하는 학습 단계, 사용자에게 적정한 아이템을 추천하는 추천 단계, 추천의 결과를 평가하는 평가 단계로 나누어진다. 3단계의 개략적인 형태는 다음 <그림 2>와 같다.



<그림 2> 웹 개인화 추천시스템 구성도

3.1 학습 단계

이 단계에서는 사용자 프로파일내의 사용자 프로파일 특성치와 이력 데이터내의 사용자별 웹 사용 데이터를 추출하여 SOM 학습을 한다. SOM은 Kohonen(1988)이 제시한 무감독 학습기법으로 입력데이터를 자기 조직화 학습 알고리즘에 따라 세분화된 N*N개의 출력노드로 클러스터링을 해주는 신경망 기법이다. 이때 각 출력노드는 세분화된 사용자 그룹을 나타낸다. 학습된 SOM을 가지고 사용자별 데이터 특성치를 이용하여 N*N개의 사용자 그룹으로 클러스터링 할 수 있다. 학습 데이터 셋 L은 사용자별(u_i)로 사례를 형성하여 구성된다.

$$L = \{u_1, u_2, \ldots, u_l\}$$

사용자별 사례는 사용자 프로파일로부터 추출한 사용자의 정적인 정보(s_{ij})와 이력 데이터로부터 추출한 아이템(상품, url, 개체 등)의 카테고리(도메인) 정보로 구성된다. 도메인 정보는 RFM 기준에 따라 각 도메인에 속한 아이템의 최근 발생 시점 정보(dR_{ik}), 빈도 수 정보(dF_{ik})와 총 구매량 정보(dM_{ik})로 나타낸다. RFM기준에서 응용분야에 따라 어느 기준은 크게 영향을 주지 못하거나 축적된 데이터의 특성상 적용이 어려운 경우가 있다. 예를 들면, R 기준은 이벤트의 발생 시점이 아주 중요한 응용이나 데이터가 몇 년 이상 축적되어 시간 흐름에 따른 영향이 큰 경우에 적용하면 효과적일 것이다. 만약에 그렇지 않은 경우에는 응용이나 데이터의 특성에 따라 R의 영향을 동일한 것으로 가정하고 F와M 기준 만을 이용하여 사용자의 데이터를 구성하는 것이 실무적인 면에서 의미가 있을 수 있다. 이 연구에서는 응용 분야의 데이터 특성에 따라 RFM 기준을 선택적으로 사용할 것을 제안한다. RFM 세 기준을 모두 반영한 경우 사용자 정보는 다음과 같이 구성된다.

$$u_i = \langle (s_{i1}, s_{i2}, \ldots, s_{im}), (dR_{i1}, dR_{i2}, \ldots, dR_{iq}), (dF_{i1}, dF_{i2}, \ldots, dF_{iq}), (dM_{i1}, dM_{i2}, \ldots, dM_{nq}) \rangle$$

사용자별 m개의 정적인 정보와 3q개의 도메인 특성치(RFM 기준을 모두 적용시에)는 하나의 사례로 표현되고 이 사례들의 집합으로 초기

학습 데이터 셋을 구성한다. 초기 학습 데이터 셋은 필요한 데이터 전처리 과정을 거쳐 학습 데이터 셋으로 변환된다. 전처리 과정에서 데이터 정제, 변환, 통합, 감축이나 확장 등의 작업이 수행된다. 특히 변환과정에서는 각 특성치의 값을 [0,1]사이의 값으로 정규화 하는데 이는 값의 크기에 따른 학습과정에서의 착시현상을 방지하기 위한 것이다. 사용자별로 구성된 학습 데이터를 SOM으로 학습시킨다.

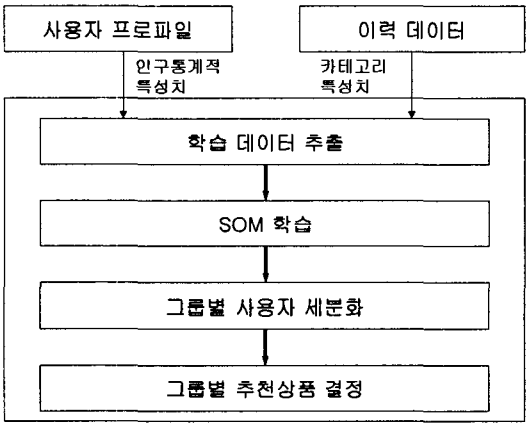
SOM을 학습하는 경우 출력 계층은 N*N의 격자형 형태가 되며, 적정한 N은 미리 결정하여 지정하여야 한다. 이 연구에서는 출력계층의 수 N*N이 도메인의 수 q보다 크게 하고, 또한 사용자의 수를 고려하여 적정한 N을 결정한다. 이때 N*N의 각 노드는 하나의 사용자 그룹을 나타낸다. 학습 후의 학습 정보는 SOM의 입력노드와 출력노드를 연결하는 가중치로 표현된다. 각 학습 사례를 학습된 SOM에 통과시키면 출력노드 N*N개의 출력값이 결정된다. 가장 큰 출력값을 갖는 노드가 대상 사례의 출력 노드가 된다. 이 과정이 끝나면 모든 학습사례는 어느 한 출력노드에 속하게 된다. 여기서 각 출력노드는 하나의 사용자 그룹을 의미한다. 각 사례는 한 사용자의 특성치를 나타내고 있으며 결과적으로 사용자들은 N*N개의 사용자 그룹으로 클러스터링 되어진다.

다음에 사용자 그룹별로 선호도가 높은 아이템을 찾는다. 아이템의 선호도를 어떤 기준으로 측정할 것인가가 고려사항이 되는 데, 이 연구에서는 RFM기준을 적용하여 아이템의 선호도를 결정하는 방안을 제안한다.

각 사용자 그룹 g에서 각 아이템(obj_i)의 선호도 점수는 RFM의 가중평균으로 구한다.

$$Score(g,obj_i)=\frac{w_r R(g,obj_i)+w_f F(g,obj_i)+w_m M(g,obj_i)}{w_r+w_f+w_m} \quad (7)$$

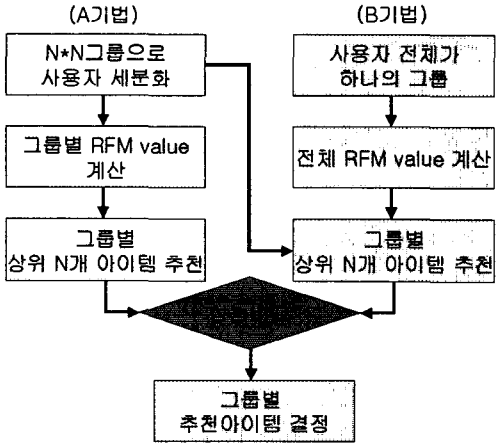
식(7)에서 $Score(g,obj_i)$는 사용자 그룹 g에서 i번째 아이템 obj_i의 선호도 점수를 나타낸다. w_r은 R의 가중치를, w_f는 F의 가중치를, w_m은 M의 가중치를 나타낸다. $R(g, obj_i)$, $F(g, obj_i)$, $M(g, obj_i)$는 사용자 그룹 g에서의 i번째 아이템(obj_i)의 RFM 값을 나타내는데, 각 RFM값은 같은 척도로 정규화한다. 응용 분야 및 데이터의 특성에 따라 RFM 기준중에서 어느 기준은 영향이 같은 것으로 가정할 수 있으며 이 경우 그 기준은 무시하고 나머지 기준을 이용하여 선호도 점수를 구하게 된다. <그림 3>은 학습 단계 절차에 대해 나타내고 있다.



<그림 3> 학습단계

사용자 그룹 별 추천 아이템의 결정은 <그림 4>와 같은 절차를 따른다. SOM을 이용하여 사용자 그룹을 클러스터링한 경우 어느 그룹내의 사용자들은 아이템의 선호도에 있어 아주 높은

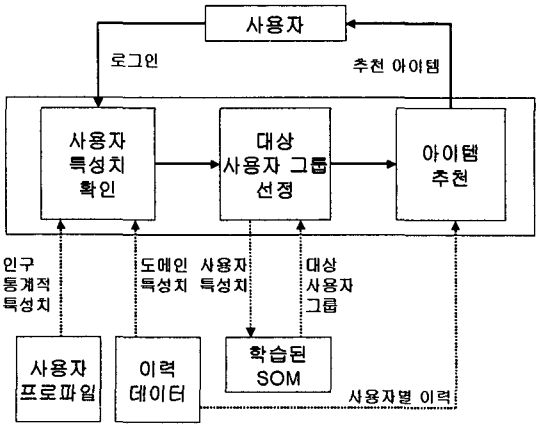
유사성을 보이는 반면에 또 다른 어느 그룹은 개성이 강한 사용자들의 특성으로 인해 같은 그룹으로 속해 있지만 아이템의 선호도가 사용자의 개성에 따라 개별적으로 작용하고 따라서 같은 그룹내 다른 사용자와의 관련성이 아주 적을 수도 있다. 이 경우 아이템의 추천은 같은 그룹에서 추천하는 것보다는 전체 사용자를 하나의 그룹으로 하여 아이템을 추천하는 것이 더 효과적일 것이다. 따라서 이 시스템에서는 대상 사용자가 속한 사용자 그룹의 선호도 정보를 이용하여 추천하는 경우(A기법)와 사용자 전체 그룹의 선호도 정보를 이용하여 추천하는 경우(B기법)의 성능을 비교하여 더 우수한 추천 성능을 나타내는 기법을 해당 사용자 그룹의 추천 아이템으로 선정한다.



<그림 4> 그룹별 추천아이템 결정 절차

3.2 추천 단계

추천 단계에서는 각 아이템에 대한 사용자 특성치 확인, 대상 사용자 그룹의 선정, 선정된 사용자 그룹에서의 선호도가 높은 상품 추천 순으로 진행이 된다(<그림 5> 참조).



<그림 5> 추천 단계

먼저 사용자 특성치 확인과정에서는 사용자 프로파일내의 인구통계학적 특성치 데이터와 이력 데이터로부터 도메인별 특성치 데이터를 추출하여 사용자 특성치 데이터를 구성한다. 만약 처음 방문한 사용자처럼 도메인별 특성치가 없는 경우에는 0과 같은 일정한 값으로 가정한다. 이 경우 사용자의 인구통계학적 특성치만을 사용하게 된다.

다음에 사용자 특성치 데이터를 이용하여 대상 사용자 그룹을 선정한다. 학습된 SOM에 사용자 특성치를 입력하면 사용자가 속하는 대상 사용자 그룹을 SOM이 출력하게 되고 학습단계에서 미리 결정한 아이템을 추천한다.

3.3 평가 단계

추천된 아이템에 대한 사용자의 반응을 분석하여 현재의 추천 시스템에 대한 평가를 한다. 추천된 아이템에 대한 사용자의 평균 평점(RA: Recommend Average) 척도와 추천되지 않은 아

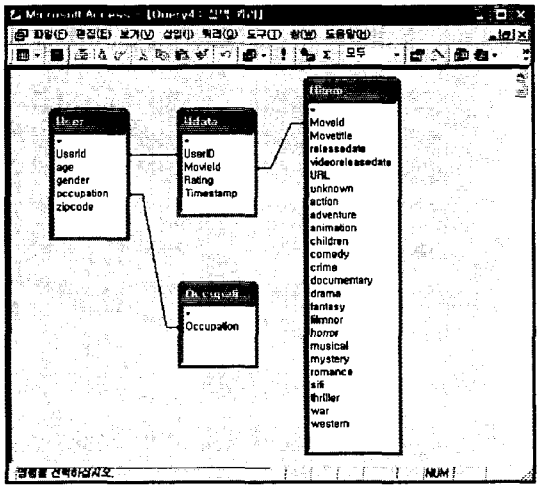
이템에 대한 평균 평점(NRA: Non-Recommend Average) 척도를 비교한다.

만약 (RA-NRA)가 정해진 값 d이상이면 현재의 추천 시스템은 효과가 있음을 의미하며 효과변수(EV: Effect Value)에 1를 증가한다. 그렇지 않으면 추천 아이템에 비해 다른 아이템들에 대한 선호도가 더 큰 것은 의미하여 이 경우 현재의 추천 시스템은 변경이 되어야 함을 나타낸다. 이 경우 비효과 변수(NEV: Non-Effect Value)에 1를 추가한다. 최근의 사용자 M명 (M = EV + NEV)에 대한 추천 시스템의 평가 결과에 따라 추천 시스템의 변경 여부를 결정한다. 즉 (NEV / M) >= c 일 때, 3.1절에서 설명한 학습을 다시 하도록 요청한다. 이 경우 SOM은 최근의 사용자 파일과 이력 데이터를 가지고 새로운 사용자 그룹을 클러스터링하게 된다.

4. 실험

4.1 MovieLens 데이터 셋

이 연구에서 제안된 개인화를 위한 협력 여과 기법에 대한 절차를 MovieLens 영화 추천시스템 데이터에 적용한다. MovieLens 데이터 셋은 미네소타 대학의 GroupLens Research Project에 의해 수집되었다. 이 데이터 셋은 1682 영화에 대해 943 사용자가 1에서 5점 사이의 값으로 점수를 매긴 100,000개의 레코드가 축적이 되어 있다. 각 사용자는 최소한 20개 이상의 영화에 대해 점수를 부여하고 있다. 사용자에 대한 간단한 정보로 나이, 성별, 직업, 우편번호 속성을 제공한다. 이 데이터 셋은 1997년 9월부터 1998년 4월까지의 7개월 동안 MovieLens 웹 사이트 (movielens.umn.edu)를 통해 수집된 것이다. MovieLens 데이터 셋의 테이블간의 관계는 <그림 6>과 같다.



<그림 6> MovieLens 데이터셋 테이블 관계

Udata 테이블은 각 영화에 대해 사용자가 점수를 부여한 100000개의 레코드를 갖고 있다. 각 영화에 대한 세부 정보는 Uitem 테이블이 갖고 있는데, 영화 장르는 Unknown을 포함하여 19개의 장르로 구분하고 있고 각 영화는 이중 하나 이상의 장르에 속하게 된다. 사용자에 대한 정보는 User 테이블에서 제공하는 데 나이, 성별, 직업, zipcode의 간단한 사용자 정보를 알려준다. Udata 테이블은 5-겹 상호 검증(5-fold cross validation)을 위해 20000개씩 랜덤하게 5개의 table로 분리하고 이중 하나는 시험용 테이블로 나머지 4개는 학습용 테이블로 구성한 5개씩의 학습용 테이블(u1base ~ u5base), 시험용 테이블(u1test ~ u5test)이 MovieLens 데이터셋에 미리 제공되어 있다. 따라서 이 연구에서는 GroupLens 연구팀에 의해 제공되고 있는 데이터

셋을 이용하여 5-겹 상호 검증 방법을 사용하여 실험하고 그 결과를 살펴보기로 한다.

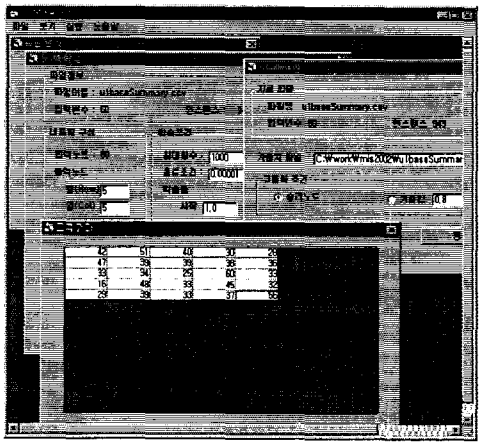
4.2 실험절차

제안된 개인화 방안의 적용 가능성 평가를 위해 MovieLens 데이터 셋을 이용한 실험절차를 이 절에서 제시한다.

■ 학습 단계

SOM 학습을 위한 학습 데이터 셋을 구성한다. 학습 데이터는 사용자의 정적인 데이터와 이력 데이터로 구성된다. 정적인 데이터는 사용자별 인구통계적 특성치인 나이, 성별, 직업 데이터를 Udata 테이블로부터 추출한다. 이력 데이터는 각 도메인 특성치로 구성되는 데 MovieLens 데이터의 경우 19개의 장르가 19개의 도메인을 구성하게 된다. 각 도메인은 3장에서 제안한 대로 RFM 기준에 따라 3개의 도메인 특성치로 확장된다. MovieLens 데이터 셋을 이용한 실험에서는 7개월 동안 수집된 데이터로 R의 영향이 크지 않으며 5겹 상호 검증을 위해 미리 제시된 데이터 셋이 레코드가 생성된 시기에 관계없이 랜덤하게 추출하여 구성하였기 때문에 R 기준은 고려하지 않고 F와 M기준만을 고려하여 실험을 실시하였다.

먼저 학습용 데이터 셋(u1base ~ u5base)을 가지고 <그림 6>의 테이블 관계를 이용하여 SOM의 학습을 위한 데이터 셋을 구성한다. SOM 학습용 데이터 셋은 943 사용자와 일대일 대응하는 943 레코드로 구성되고, 각 레코드는 인구통계적 특성치와 19개의 도메인을 F와 M으로 확장한 38개의 도메인 특성치로 구성된다. 직업 특성치는 각 직업을 하나의 속성으로 확장하여 표시한다(주어진 데이터에서 직업의 수가 20개임으로 20개의 속성으로 확장하고 사용자가 해당 직업을 가지면 1 아니면 0을 표시한다). 성별은 남자의 경우 1을 여자의 경우 0으로 나타낸다. F에 대한 도메인 특성치는 각 사용자가 본 도메인에 해당하는 영화의 수를, M에 대한 도메인 특성치는 영화에 대해 사용자가 1에서 5점 사이의 평점을 부여하는 데 도메인에 해당하는 영화의 평균 평점을 나타낸다. 모든 특성치는 0에서 1사이의 값을 갖지 않는 특성치의 경우 정규분포를 이용하여 정규화한다(SOM은 유클리디안 거리척도에 의해 학습하기 때문에 학습 데이터는 같은 스케일로 정규화하여야 한다). 이 과정을 거치면 사용자별로 60개의 속성을 갖는 943 레코드를 생성하며 이 데이터를 이용하여 SOM을 학습한다. SOM의 출력계층의 노드는 도메인의 19보다 큰 25(5*5)개로 설정한다. SOM의 구현은 Visual Basic 6.0으로 구현하며 내부 알고리즘은 2절에서 제시한 절차를 따른다. <그림 7>은 SOM을 이용한 학습 결과 각 사용자를 25개의 그룹으로 클러스터링한 예를 보여준다.



<그림 7> SOM을 이용한 그룹화 예

SOM의 학습결과 각 사용자는 25개의 사용자 그룹 중 한 그룹에 속하게 된다. 다음에 사용자 그룹별 선호도가 높은 영화를 찾고 추천할 영화를 결정한다. 각 사용자 그룹별 영화의 Score는 다음 기준에 의해 계산한다.

$F(g, obj_i)$는 사용자 그룹 g에서의 i번째 아이템에 대한 빈도(Frequency)와 관련된 점수이다. 빈도를 고려할 때에는 한 아이템에 대한 좋은 평가의 빈도와 나쁜 평가의 빈도를 동시에 고려하여야 한다. 따라서 F값은 다음 식 (8)과 같이 정의한다.

$$F(g,obj_i)=\max\{F_a(g,obj_i)-F_b(g,obj_i),\ 0\} \qquad (8)$$

식 (8)에서 $F_a(g,i)$는 사용자 그룹 g에 속하는 사용자들이 영화 i에 대해 평점을 4이상 준 빈도수, $F_b(g,i)$는 사용자 그룹 g에 속하는 사용자들이 영화 i에 대해 평점을 3이하 준 빈도수를 나타낸다. 식 (8)에서 구한 $F(g, obj_i)$는 각 사용자 그룹내에서 정규분포를 이용하여 0에서 5점 사이로 정규화한다.

$M(g, obj_i)$값은 사용자 그룹 g에 속하는 사용자들이 영화 i에 부여한 평균 평점으로 정의한다. 평점은 1에서 5점 사이에서 사용자가 부여했기 때문에 $M(g, obj_i)$의 값은 0에서 5점 사이에 분포하게 된다. F와 M값은 [0,5]의 범위내의 값을 가지며 두 기준의 가중치는 같은 것으로, 즉 각각 0.5로 정의한다.

각 사용자 그룹내의 영화에 대한 선호도는 식 (7)을 이용해 계산한다. 이 실험에서는 실험데이터의 한계상 R기준을 고려하지 않기로 하였으므로 w_r=0이다. 각 사용자 그룹별로 선호도 점수가 높은 영화 순서로 정렬하여 저장한다.

■ 추천 단계

사용자가 로그인하면 사용자의 인구통계적 특성치와 장르별 영화 정보 그리고 정규화 정보를 가지고 SOM의 입력 데이터 셋을 구성하고 학습된 SOM을 통해 대상 사용자 그룹을 결정한다. 다음에 로그인한 사용자가 보지 않은 영화에 대해 선호도 점수가 높은 영화 순으로 추천을 한다.

다음 <표 1>은 MovieLens 데이터셋에서 미리 주어진 5개의 시험용 데이터셋(u1test~u5test)을 이용한 5겹 상호검증 실험의 결과이다.

<표 1>의 MAE는 사용자 그룹별로 상위 N개의 영화를 추천한 경우 영화의 선호도 값과 실제 사용자가 부여한 평점사이의 차이에 대한 평균으로 MAE는 다음과 같이 계산한다(Sarwar et al., 2001).

$$MAE=\frac{\sum_{i=1}^{Ng}|v_i-q_i|}{N_g} \qquad (9)$$

Ng는 사용자 그룹 g의 상위 추천 영화의 개수를, v_i는 사용자 그룹별 영화 i에 대한 선호도 점수를, q_i는 사용자 그룹 g에 속하는 사용자가 영화 i에 대해 실제로 부여한 점수를 의미한다. 사용자가 로그인한 경우 사용자가 속한 대상 사용자 그룹의 선호도 순으로 영화 10개를 추천했을 때의 5개 시험용 데이터 셋의 평균 MAE는 0.668임을 <표 1>에서 알 수 있다. MAE의 값은 식 (9)에서 알 수 있듯이 적을수록 추천 성능이 좋음을 의미하는데, 추천 영화의 수가 증가함에 따라 MAE 기준에서 보면 추천 성능이 떨어짐을 보여주고 있다. DSA는 사용자의 선택을 얼마나 잘 도와주는 가의 효과성을 평가하는 척도이다(Sarwar et al,. 2001). 이 실험에서는 로그인한

<표 1> 5겹 상호검증 실험결과

		N=10	N=20	N=30	N=40	N=50
MAE	U1	0.638	0.706	0.744	0.762	0.765
	U2	0.672	0.717	0.745	0.757	0.774
	U3	0.714	0.749	0.767	0.792	0.810
	U4	0.642	0.722	0.749	0.770	0.790
	U5	0.674	0.716	0.751	0.759	0.783
평균MAE		**0.668**	**0.722**	**0.751**	**0.768**	**0.784**
DSA	U1	0.880	0.845	0.822	0.807	0.801
	U2	0.866	0.828	0.806	0.789	0.770
	U3	0.839	0.817	0.797	0.781	0.759
	U4	0.884	0.841	0.816	0.796	0.780
	U5	0.855	0.833	0.808	0.795	0.774
평균 DSA		**0.865**	**0.833**	**0.810**	**0.794**	**0.777**

* MAE(Mean Absolute Error), DSA(Decision Support Accuracy)

사용자에 대해 추천한 영화를 사용자가 4이상의 평점을 부여한 경우에 효과적으로 추천한 것으로 가정하였고, 이때 DSA는 1을 사용자가 3이하를 부여한 경우에는 DSA는 0으로 정의하였다. <표 1>에서 보면 사용자 그룹별로 영화 10개를 추천한 경우 5개 시험용 데이터 셋에 적용한 결과 평균 DSA가 0.865인데, 이는 추천한 영화의 86.5%가 사용자로부터 4점 이상의 평점을 부여 받았음을 의미한다. DSA는 그 값이 높을수록 추천 성능이 우수함을 의미하는 데, MAE와 마찬가지로 추천 영화의 수가 증가하게 되면 추천 성능이 떨어짐을 알 수 있다. 따라서 많은 영화를 추천하기 보다는 상위 20개 이내의 영화 중에서 사용자가 보지 않은 영화를 추천하는 것이 좋을 것으로 판단된다.

■ 평가 단계

영화 추천후의 시스템에 대한 평가는 사용자가 부여한 추천한 영화에 대한 평균 평점(RA) 척도와 비 추천 영화에 대한 평균 평점(NRA) 척도를 비교하여 판정한다. 사용자별로 추천 영화의 평균 평점이 비추천 영화의 평균 평점보다 일정 값 d이상인 경우 추천 효과변수(EV)에 1을 증가시키고 그렇지 않은 경우 비효과변수(NEV)에 1을 증가시킨 다음, 식 (NEV/ (EV+NEV))의 결과가 일정 값 c 이상일 때 현재의 추천 시스템을 수정한다. 수정 시에는 새롭게 축적된 데이터를 포함하여 SOM을 재학습하고 새로운 사용자 그룹으로 클러스터링 한다. 평가 단계의 실험은 영화를 추천하고 추천한 영화에 대해 사용자가 반응한 결과에 대해 실험을 실시해야 하나 현재 주어진 MovieLens는 이 연구에서 의도하는 그러한 과정이 반영되지 않은 즉 다른 목적 하에 축적한 데이터로 평가단계의 실험은 거의 불가능하다. 따라서 평가 단계의 실험은 다음 <표 2>의 예로 대신한다. 여기서 RA와 NRA에 대한 척도로 DSA를 사용하기로 한다.

5번째 사용자에 대한 평가 결과 시스템의 V값

<표 2> 평가 단계의 예(d=0.2, V=NEV/(EV+NEV), c= 0.4 일 때)

user	RA	NRA	EV	NEV	V	수정
1	0.9	0.7	1	0	0	n
2	0.8	0.6	2	0	0	n
3	0.7	0.6	2	1	0.33	n
4	0.8	0.5	3	1	0.25	n
5	0.8	0.7	3	2	0.4	y

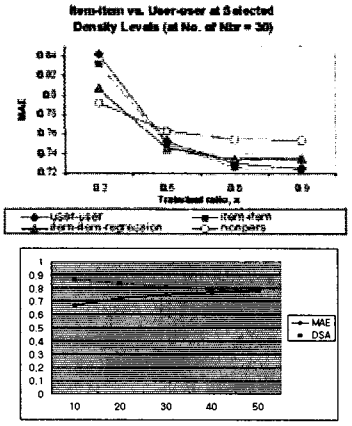
이 0.4가 되어 미리 정한 c의 값 이상임으로 현재의 시스템을 수정하도록 한다. 파라미터 c와 d는 응용 분야에 적합하도록 설정하도록 한다.

4.3 토의

본 연구는 협력 여과 기법에서 주로 사용하는 유사성 척도가 사용자 집단이 커짐에 계산의 복잡도가 지수적으로 증가하는 문제를 해결하기 위한 방안을 제시하는 것이 주 목적이다. 먼저 SOM을 이용하여 사용자를 세분화하고 로그인한 사용자에 대해서 대상 사용자 그룹을 확인한 다음 대상 사용자 그룹의 선호도가 높은 아이템을 추천하는 절차를 제안하고 있다. 선호도 점수는 FM 기준을 이용하여 부여하고 있는데, 사용자 그룹의 세분화와 사용자 그룹별 선호도 점수는 오프 라인에서 미리 계산하여 저장함으로써 추천 시스템 운용시의 연산 복잡도를 일정하게 유지시킬 수 있게 된다.

<그림 8>은 연구에서 제안한 시스템과 MovieLens 데이터셋을 제공한 GroupLens 연구팀의 최근 실험결과(Sarwar et al., 2001)를 비교한 것이다. GroupLens 연구팀은 협력적 여과 기법을 이용한 추천시스템 연구의 선구자라 할 수 있으며 많은 연구결과를 발표하고 있다. 따라서 MovieLens 데이터셋을 이용한 연구에서는 최근에 발표한 Sarwar등의 실험결과와 비교함으로써 성능을 어느 정도 평가하는 것이 가능하다. 제안한 시스템을 MovieLens 데이터 셋에 적용해 본 결과 사용자 그룹별 상위 10개의 영화를 추천한 경우 MAE가 0.668로 나타났다. Sarwar등(2001)은 실험에서 80000개의 학습용 데이터와 20000개의 시험 데이터를 적용한 경우 (train/test ratio x=0.8 일때), 그리고 가장 추천 성능이 좋은 것으로 제시하고 있는 이웃의 크기는 30일 때의 네 가지 기법의 성능을 비교하고 있다. <그림 8>의 범례에서 user-user는 전통적 협력적 여과 기법을, item-item은 협력적 여과 기법의 규모의 문제를 해결하기 위해 item의 유사성을 이용하는 기법을, item-item-regression은 item-item기법에서 사용자에게 item을 추천하기 위한 스코어 계산에서 사용자가 유사 item들에 부여한 평점(rating)을 직접 사용하는 데 비해 평점 계산을 선형 회귀식을 이용한 경우의 기법을, nonpers는 Herlocker등(1999)이 제안한 비개인화(naïve nonpersonalized) 알고리즘을 나타낸다. x=0.8일 때 item-item 기법이 약 0.725로 규모의 문제를 해결하면서 또한 전통적 협력적 여과 기법인 user-uesr기법에 비해서도 성능이 나음을 보여주고 있다. 그러나 2.1절에서 지적한 것처럼 item-item 기법은 협력적 필터링 기법의 규모의 문제를 해결하기 위해 내용 기반 여과 기법을 사

용한 것으로 내용 기반 여과 기법의 한계를 가지게 된다. 본 연구에서 제안한 방안과 비교해 보면, 동일 조건 실험인 x=0.8에서 추천 영화의 수가 20개 이내인 경우 Sarwar et al.(2001)의 결과에 비해 제안된 방안이 더 나은 추천 성능을 보임을 알 수 있다. 따라서 Sarwar et al.(2001)의 결과와 비교해 봤을 때 제안된 시스템은 추천 성능면에서도 받아들일 만한 것임을 알 수 있다.



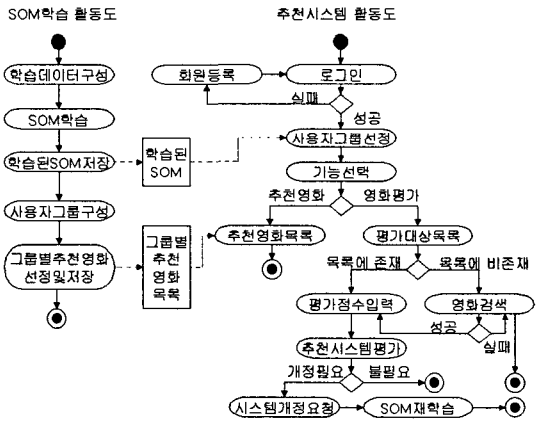
Grouplens Research Team의 성능
[Sarwar et al., 2001]
-neighborhood size=30
-Train dataset: 80000
-Test dataset: 20000
-10-fold cross validation
- MAE:
item-item: 약 0.725
item-item-r: 약 0.735
user-user: 약 0.73

본 연구의 성능
-N=10일 때
MAE=0.668
-N=20일 때
MAE=0.722

<그림 8> 본 연구와 Sarwar(2001)연구의 실험 결과 비교

제안된 시스템은 또한 협력적 여과 기법이 가지는 몇 가지 문제점에 대해서도 완화시킬 수 있다. 첫번째 새로운 사용자의 경우 이력 데이터가 없으므로 기존의 방법에서는 유사 사용자의 선정에 어려움을 가졌다. 이 경우 제안 시스템은 새로운 사용자의 인구통계적 특성치만을 이용하여 사용자 그룹을 선정하는 것이 가능하다. 두번째 새로운 아이템이 발생한 경우 새로운 아이템의 사용 이력이 없으므로 추천에서 빠지는 문제가 있다. 그러나 제안한 방안에서는 신규 아이템의 경우 전체 사용자를 하나의 그룹으로 하여 RFM 선호도 점수로 추천을 하다가(B기법에 의해) 어느 정도 데이터가 쌓이게 되면 A기법과 B기법의 비교에 의해 추천하는 것이 가능해진다.

제안된 추천 시스템의 프로토타입을 활동도(Activity Diagram)로 표현해 보면 다음 <그림 9>와 같다. <그림 9>에서 점선은 활동 사이의 입출력을 표현하고 있으며, SOM학습 활동도와 추천시스템 활동도는 별개의 도표로 그려지는데, 여기에서는 입출력 데이터의 연관성을 보여주기 위해 연결하여 표현하고 있다.



<그림 9> 상품추천시스템 프로토타입에 대한 활동도

5. 결론

개인화된 정보를 제공하기 위한 협력 여과 기법에 대한 많은 연구가 이루어지고 있는 데, 유사 사용자들을 찾는 과정에서 모든 사용자와의 유사성을 계산하여 구하는 것은 사용자의 수가 점차 많아지게 되는 웹의 특성상 계산의 복잡도가 지수적으로 증가하게 된다. 또한 많은 연구가

추천 그 자체에서 그치고 있고 추천의 결과 평가를 통한 추천 시스템의 변경 시점에 대한 연구는 거의 이루어지고 있지 않다.

이 연구에서는 신경망의 SOM 기법을 이용하여 성향이 유사한 사용자 그룹으로 미리 클러스터링하고 사용자 그룹별로 RFM 기준에 의해 각 아이템의 선호도 점수를 계산한 다음, 사용자 그룹별 추천 아이템을 미리 결정하여 아이템을 추천하는 절차를 제안하였다. SOM 학습은 오프라인에서 이루어지며 추천 단계에서는 학습된 SOM을 이용하기 때문에 사용자 집단이 지속적으로 증가하더라도 추천 시스템의 운용시의 계산의 복잡도를 일정하게 유지할 수 있으며 추천 성능도 이전의 유사성 척도를 이용한 경우보다 비교적 좋은 것을 알 수 있다. 또한 추천 결과에 대한 평가를 통해 추천 시스템의 개정 시점을 알려주는 방안에 대해서도 제안하고 있다. 제안된 절차는 개인화 시스템의 협력적 여과 기법에 효과적으로 적용됨으로써 관계마케팅을 위한 좋은 도구가 될 수 있을 것이다.

그러나 제안한 기법은 RFM기준의 사용을 제안하였지만 실험데이터의 한계상 R의 적용이 쉽지 않았고 따라서 제안한 기법은 다른 데이터 셋에 R을 고려한 다양한 추가적인 실험을 진행할 필요가 있다. 추천 성능의 개선을 위한 관점에서 SOM의 입력 데이터 셋의 구성에서 데이터 전처리 과정을 통해 결과를 좋게 할 수 있도록 주요 변수만을 추출하여 학습용 입력 데이터 셋을 구성하는 절차 등에 대한 연구도 필요하다. 또한 다른 기법과의 하이브리드 결합을 통한 성능 개선 방안 등 개인화 시스템의 성능 개선을 위한 여러 방안이 향후 과제로서 계속 진행되어져야 할 것이다.

참고문헌

안현철, "데이터 마이닝을 활용한 인터넷 쇼핑몰의 상품 추천 시스템 개발", 한국과학기술원 석사학위논문, 2002.

Bigus, J.P. and J. Bigus, *Constructing intelligent agents using JAVA*, John Willy & Son, New York, 2001.

Breese, J.S., D. Heckerman and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, 1998, 43-52.

Bult, J.R. and T. Wansbeek, "Optimal selection for direct mail," *Marketing Science,* Vol.14, No.4(1995), 378-394.

Changchien, S.W. and T.C. Lu, "Mining Association Rules Procedure to Support On-line Recommendation by Customers and Products Fragmentation," *Expert Systems with Applications,* Vol.20(2001), 325-335.

Cohen, W.W. and W. Fan, "Web-collaborative filtering: recommending music by crawling the Web," *Computer Networks,* Vol.33(2000), 635-698.

Goldberg, D., D. Nichols, B.M. Oki and D. Terry, "Using Collaborative Filtering to Weave an Information Tapestry," *Communications of the ACM*, Dec 1992.

Gupta, D and K. Goldberg, "Jester 2.0: A Linear Time Collaborative Filtering Algorithm Applied to Jokes," In *Proc. of the ACM SIGIR '99*, 1999.

Ha, S.H. and S.C. Park, "Application of Data Mining Tools to Hotel Data Mart on the Intranet for Database Marketing," *Expert Systems with Applications,* Vol.15 (1998), 1-31.

Herlocker, J., J. Konstan, A. Borchers and J. Riedl, "An Algorithm Framework for Performing Collaborative Filtering," *Proceedings of the*

1999 Conference on Research and Development in Information Retrieval, Aug 1999.

Hill, W., L. Stead, M. Rosenstein and G. Furnas, "Recommending and evaluating choices in a virtual community of use," *Proceedings of ACM CHI '95 Conference on Human Factors in Computing Systems*, 1995, 194-201.

Kohonen, T., "An introduction to neural computing," *Neural Networks*, 1, 1988, 3-16.

Kohrs, A. and B. Merialdo, "Creating user-adapted Websites by the use of collaborative filtering," *Interacting Computers*, Vol.13 (2001), 695-716.

Konstan, J., B. Miller, D. Maltz, J. Herlocker, L. Gordon and J. Riedl, "GroupLens: Applying Collaborative Filtering to Usenet News," *Communications of the ACM*, Vol.40, No.3 (1997), 77-87.

Lee, C.H., Y.H. Kim and P.K. Rhee, "Web personalization expert with combining collaborative filtering and association rule mining technique," *Expert Systems with Applications*, Vol.21(2001), 131-137.

Lee, D.S., G.Y. Kim and H.I. Choi, "A Web-based collaborative filtering system," *Pattern Recognition*, Vol.36, No.2(Feb 2003), 519-526.

Resnick, P. and H.R. Varian, "Recommender Systems." *Special issue of Communications of the ACM*, Vol.40, No.3(1997).

Sarwar, B., G. Karypis, J. Konstan and J. Riedl, "Application of Dimensionality Reduction in Recommender System-A Case Study," *ACM WebKDD 2000 Web Mining for E-Commerce Workshop*, 2000

Sarwar, B., G. Karypis, J. Konstan and J. Riedl, "Item-Based Collaborative Filtering Recommendation Algorithms," *Proceedings of the 10th International World Wide Web Conference*, Hong Kong, May 2001, 285-295.

Sarwar, B., G. Karypis, J. Konstan and J. Riedl, "Recommender Systems for Large-scale E-Commerce: Scalable Neighborhood Formation Using Clustering," *Proc. of the Fifth International Conference on Computer and Information Technology (ICCIT 2002)*, 2002.

Shardanand, U. and P. Maes, "Social Information Filtering: Algorithms for automating 'word of mouth'," *Proceedings of ACM CHI '95 Conference on Human Factors in Computing Systems*, 1995, 210-217.

Smith, K.A. and A. Ng, "Web Page Clustering using Self-Organizing Map of User Navigation Patterns", *Decision Support Systems*, Vol.35(2003), 245-256.

Ungar, L.H. and D.P. Forster, "Clustering Methods for Collaborative Filtering," In *Workshop on Recommender Systems at the 15th National Conference on Artificial Intelligence*, 1998.

Abstract

Collaborative Filtering System using Self-Organizing Map for Web Personalization

Boo-Sik Kang*

This study is to propose a procedure solving scale problem of traditional collaborative filtering (CF) approach. The CF approach generally uses some similarity measures like correlation coefficient. So, as the user of the Website increases, the complexity of computation increases exponentially. To solve the scale problem, this study suggests a clustering model-based approach using Self-Organizing Map (SOM) and RFM (Recency, Frequency, Momentary) method.

SOM clusters users into some user groups. The preference score of each item in a group is computed using RFM method. The items are sorted and stored in their preference score order. If an active user logins in the system, SOM determines a user group according to the user's characteristics. And the system recommends items to the user using the stored information for the group. If the user evaluates the recommended items, the system determines whether it will be updated or not.

Experimental results applied to MovieLens dataset show that the proposed method outperforms than the traditional CF method comparatively in the recommendation performance and the computation complexity.

Key words : Collaborative Filtering, Self-Organizing Map, Web Personalization

* Dept. of Business Administration and information, Mokwon University