지능정보연구 제16권 제4호 2010년 12월(pp.159~172)

사회연결망분석과 인공신경망을 이용한 추천시스템 성능 예측

조윤호
국민대학교 경영대학 경영정보학부
(www4u@kookmin.ac.kr)

김인환
국민대학교 대학원 e-비즈니스학과
(inhwan@kookmin.ac.kr)

협업필터링 추천은 다양한 분야에서 활용되고 있지만 트랜잭션 데이터의 성격에 따라 추천 성능에 현저한 차이를 보이고 있다. 기존 연구에서는 이러한 추천 성능의 차이가 나타나는 이유에 대한 설명을 구체적으로 제시하지 못하고 있고 이에 따라 추천 성능의 예측 또한 연구된 바가 없다. 본 연구는 사회네트워크분석과 인공신경망 모형을 이용하여 협업필터링 추천시스템의 성능을 예측하고자 한다. 본 연구의 목적을 달성하기 위해 국내 백화점의 트랜잭션 데이터를 기반으로 형성되는 고객간 사회 네트워크의 구조적 지표를 측정한 후 이를 기반으로 인공신경망 모형을 구축하고 검증한다. 본 연구는 협업필터링 추천 성능을 예측할 수 있는 새로운 모형을 제시하였다는 점에서 그 의의가 있으며 이를 통해 기업들의 협업필터링 추천시스템 도입에 대한 의사결정에 도움을 줄 수 있을 것으로 기대된다.

논문접수일 : 2010년 11월 21일 게재확정일 : 2010년 12월 06일 교신저자 : 조윤호

1. 서 론

추천시스템은 통계적 기법과 지식탐사기술을 사용하여 고객 개개인의 취향에 가장 부합하는 상품 또는 서비스를 추천해주는 시스템으로서, 고객들의 구매 편의를 도모하고 교차판매 및 매출 증대에 초점을 맞춘 시스템이다(조윤호, 방정혜, 2009). 현재까지 추천시스템을 구현하기 위한 다양한 기법들이 개발되어 왔는데, 이중에서 협업필터링이 가장 성공적인 추천기법으로 알려져 있으며 Amazone.com, Netflix.com, CDNow.com 등 수많은 기업들이 협업필터링을 통해 고객에게 추천서비스를 제공하고 있다(박종학 외, 2008; Su and Khoshgoftaar, 2009).

그럼에도 불구하고, 협업필터링의 추천 성능은 적용하는 기업의 비즈니스 형태나 소유하고 있는 정보의 특성에 따라 다르게 나타난다(Huang and Zeng, 2005; Huang et al., 2007). 그 한 예로서, 음악이나 영화 사이트에서는 추천의 정확도가 높지만, 상대적으로 e-쇼핑몰에서는 추천의 정확도가 낮은 것으로 보고되고 있다(Sarwar et al., 2000). 기업에서 협업필터링 추천시스템을 구축하려면 상당한 시간과 비용이 소요되기 때문에 구축된 추천시스템의 성과가 높지 않다면 기업 자원의 낭비를 초래할 뿐만 아니라 부정확한 추천서비스를 받는 고객들의 불만을 사게 될 것이다.

본 연구에서는 협업필터링 추천시스템을 도입하고자 하는 기업들이 사전에 추천 성능을 효율적으로 예측할 수 있는 모형을 개발하고자 한다. 예측모형의 독립변수를 선정하기 위해 사회네트워

* 본 논문은 2009년도 국민대학교 교내연구비를 지원받아 수행된 연구임.

크 분석(Social Network Analysis : SNA)을 활용한다. 사회네트워크 분석은 의사소통 집단 내 개체의 상호작용에 관심을 두고, 개체간 연결 상태 및 연결 구조의 특성을 계량적으로 파악하여 시각적으로 표현하는 분석기법이다(Wasserman, 1994; 손동원, 2002; 김용학, 2003). 협업필터링에서는 고객들 간의 상품 선호도 또는 구매 연관성을 분석한 후 유사한 고객들을 묶어 상품을 추천하는 관계를 형성하는데, 이 관계를 그래프로 나타내어 네트워크를 형성하면 이 네트워크는 사회네트워크가 된다(Ryu et al., 2006). 즉, 협업필터링 추천시스템은 고객의 구매데이터를 분석하여 사회네트워크를 인위적으로 생성한 후에 링크로 연결된 고객들(이웃고객)의 구매정보를 이용하여 특정 고객에게 상품을 추천할 수 있게 만든 일종의 사회네트워크 시스템이라고 할 수 있다.

이렇게 구축된 사회네트워크로부터 다양한 구조적 지표들을 측정할 수 있는데 이중에서 적절한 지표들을 예측 모형의 독립변수로 선정한다. 또한 별도의 협업필터링 시스템을 구축하고 추천 정확도를 계산한 후 이를 종속변수로 사용한다. 모형을 구축하기 위한 기법으로 회귀분석과 인공신경망 기법을 모두 활용할 수 있지만, 본 연구에서는 다양한 분야에서 보다 우수한 예측능력을 보이는 인공신경망(Artificial Neural Networks : ANN) 기법을 활용한다.

2. 관련 연구

2.1 협업필터링

협업필터링 추천시스템은 상품을 추천하고자 하는 고객과 취향이 유사한 고객들의 의견을 반영하여 추천 대상 고객이 아직 구매하지 않은 상품에 대한 선호도를 예측한 후 선호도가 높을 것으로 예측되는 상품을 추천해주는 시스템이다. 일반적으로 협업필터링 기반 추천 프로세스는 크게 입력 데이터 구성, 이웃 집단 탐색, 추천 상품 결정 단계로 구성된다(Sarwar et al., 2000).

그러나 협업필터링에 기반한 상품추천시스템은 다음과 같은 네 가지 근본적인 문제점을 갖고 있다(Sarwar et al., 2000; Cho and Kim, 2004; Adomavicius and Tuzhilin, 2005; 조영빈, 조윤호, 2007). 첫째는 희박성(Sparsity)의 문제이다. 협업필터링 추천시스템은 고객의 선호도 데이터를 많이 확보할수록 추천의 정확도가 높아진다. 그러나 취급하는 상품이 많아질수록 고객의 직접 평가나 구매정보 분석을 통하여 수집되는 선호도 데이터가 존재하지 않은 상품의 개수가 상대적으로 많아진다. 따라서 고객-상품 행렬은 희박 행렬(Sparse Matrix)일 수밖에 없으며, 유사집단을 탐색하는 과정에서 아주 적은 수의 선호도 데이터를 사용하므로 고객들 간의 유사도 측정 시 신뢰성이 떨어지게 된다. 이러한 현상은 결국 추천결과의 정확도를 떨어뜨리게 하는 주요인으로 작용한다. 둘째는 추천 알고리즘의 확장성(Scalability) 문제이다. 협업필터링 상품추천 프로세스에서 유사 집단 탐색과정은 일종의 lazy learning과 유사하다. 따라서 고객과 상품의 수가 증가하면 고차원화된 고객-상품 행렬로부터 유사 고객을 찾기 위한 연산량은 기하급수적으로 늘어 날 수밖에 없기 때문에 실시간 추천을 목적으로 하는 추천시스템은 심각한 시스템 확장성 문제에 직면하게 된다. 셋째는 신규고객의 상품추천 문제이다. 협업필터링 추천시스템에서 고객이 상품을 추천 받으려면 구매이력이나 상품에 대한 선호도 정보가 필요하다. 그러나 처음 방문하는 신규고객에게는 상품에 대한 선호도 정보나 상품에 대한 구매이력이 전혀 없으므로, 신규고객에게 상품을 추천하는 것이 근본적으로 어렵다. 넷째는

신상품 추천문제로서, 신상품의 경우에는 과거에 어떠한 고객도 구매한 적이 없어 이에 대한 선호도를 예측할 수 없으므로 추천이 불가능하게 된다.

이와 같은 협업필터링의 한계점을 극복하기 위하여 많은 연구자들이 추천 대상 고객이 선호하는 상품과 유사한 특성을 가진 상품을 추천하는 내용기반 필터링(Content-based Filtering)을 협업필터링을 결합한 다양한 하이브리드 추천방법(Melville et al., 2001; Cho and Kim, 2004; 김재경 외, 2005)을 제안하였고, 최근에는 사회네트워크분석을 적용한 기법들이 다양하게 논의되고 있다(박종학 외, 2009; 조윤호, 방정혜, 2009; Liu and Lee, 2010). 이러한 연구들은 협업필터링의 문제점을 개선하고 성능을 향상시키는데 초점을 맞추어왔고 많은 연구 성과가 있었지만, 추천 성능을 사전에 예측하고자 하는 연구는 현재까지 전무한 실정이다.

2.2 사회네트워크 분석

지난 40년에 걸친 사회네트워크 분야의 연구들은 참여자들의 연결패턴으로 표현되는 네트워크 구조를 통해서 사회적인 상호작용을 설명하고 이해하는 것을 추구해왔다. 즉, 사회네트워크는 팀이나 조직, 산업 등과 같은 다양한 사회적 구조의 행동을 설명하고 이해하는데 폭넓게 사용되어 왔다(Kukkonen et al., 2010). 이러한 사회네트워크 분석에서는 데이터를 매트릭스로 표현한다. 행과 열의 개체간 관계가 존재하면 1, 그렇지 않은 경우에는 0으로 입력하거나 계량값으로 표현한다(손동원, 2002).

고객과 상품의 구매관계를 구매(1), 비구매(0)로 표현한 고객-상품 매트릭스에서 고객 사이에 직접적인 관계가 없어도 인위적인 관계를 사용하여 나타낼 수 있는데 이러한 네트워크를 준연결망(Quasi network)라고 한다(김용학, 2003). 예를 들어, 동일한 제품을 1개 이상 구매한 고객들을 인위적으로 연결하는 방식으로 고객간 네트워크를 구성할 수 있다. 일반적으로 고객의 상호 관계를 나타내는 방법으로는 동일 제품 구매 빈도, 코사인 벡터, 상관계수, Jaccard 유사도 등이 이용되고 있다(김용학, 2003).

사회 네트워크는 분석의 초점을 어디에 두는가에 따라 에고 네트워크(Ego-centric network), 양자 네트워크(Dyadic network), 전체 네트워크(Total network)로 구분할 수 있다(손동원, 2002; Weare et al., 2007). 에고 네트워크는 네트워크의 한 구성원을 중심으로 그 구성원과 다른 구성원 간의 연결을 표현한 네트워크이고, 양자 네트워크는 네트워크 내의 두 구성원을 중심으로 그 사이의 연결을 표현하는 네트워크이다. 전체 네트워크는 보편적인 네트워크로 N명의 전체 행위자들로 구성된 전체 네트워크를 말한다. 또한 사회네트워크는 관계의 방향성의 유무에 따라 방향성(Directed) 네트워크와 비방향성(Undirected) 네트워크로 분류된다(Wasserman and Faust, 1994). 방향성 네트워크는 시작 점과 끝점이 존재하는 방향을 가진 관계를 포함하는 네트워크이고, 비방향성 네트워크는 두 구성원 사이의 관계가 상호 동일한 네트워크를 말한다.

기존 연구에서는 사회네트워크의 구조를 측정하기 위한 다양한 개념들이 개발되어 왔다. 먼저 네트워크의 결속(Cohesion)을 측정하기 위한 지표로는 포괄성(Inclusiveness), 연결정도(Degree), 밀도(Density), 군집화계수(Clustering coefficient), 이행성(Transitivity) 등이 있다(Frank and Harary, 1982; Wasserman and Faust, 1994; Watts, 1999; 손동원, 2002; 김용학, 2003). 사회네트워크를 구성하는 하부 집단을 규명하기 위한 측정지표로는 파

당(Faction), 결속집단(Clique) 등이 있다(Seidman and Foster, 1978; Amorim et al., 1992). 사회네트워크에서의 역할과 위치에 대한 구조적 등위성을 측정하기 위한 지표로 CONCOR, STRUCTURE 등이 개발되었다(Breiger et al., 1975; Burt 1991; 손동원, 2002). 한편 네트워크에서 중심에 위치하는 정도와 중심에 집중된 정도를 나타내는 측정지표로는 연결정도 중심성(Degree centrality), 근접중심성(Closeness centrality), 매개 중심성(Betweeness centrality), 집중도(Centralization) 등이 있다(Freeman, 1979; Bonachich, 1987; 손동원, 2002). 이 외에도 에고 네트워크에서 구조적 틈새를 위한 측정지표로는 효율성(Efficiency), 효율적 크기(Effsize), 계층(Hierarchy)등이 있고(Burt, 1992), 조직의 하향식 트리 구조를 규명하기 위해 Krackhardt(1994)는 계층(Hierarchy), 연결성(Connectedness), 효율성(Efficiency), LUB 등을 제안하였다.

앞서 기술한 준연결망의 예처럼 고객 사이의 구매 관계는 비방향성 관계를 형성하기 때문에 본 연구에서는 비방향성 전체 네트워크만을 고려하고자 한다. 따라서 위에서 열거한 사회네트워크의 구조적 지표들 중에서 에고 네트워크 수준의 측정지표와 하부구조에 관련된 측정지표들을 제외하고 비방향성 그래프에 적용 가능한 지표들만을, 즉, 포괄성, 밀도, 군집화 계수, 연결정도 집중도, 매개 집중도, 근접 집중도, 효율성을 예측모형의 독립변수로 사용하고자 한다.

3. 실험 설계

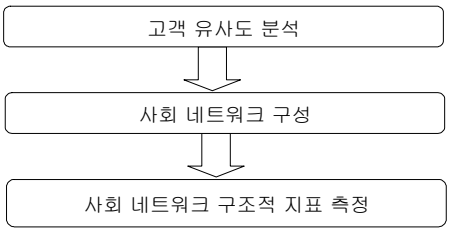
3.1 데이터 수집

추천 성능을 예측하기 위해 국내 유명 백화점 중의 하나인 H백화점의 구매 데이터를 이용하였다. 2000년 5월 1일부터 2001년 4월 30일까지 1년 동안 50,000명의 고객들이 4,038개의 상품에 대해 일으킨 1,660,814건의 구매 트랜잭션이 실험에 사용되었다. 다양한 형태와 구조적 특성을 갖는 사회네트워크를 생성하기 위해, 먼저 구매 트랜잭션을 월별로 분할하고 이를 2개월씩 조합하여 66($_{12}C_2$)개의 기본 표본을 추출했다. 이러한 66개의 표본 각각에서 무작위로 100, 200, 300명의 고객을 추출하여 표본의 수를 198개로 증대시켰다. 마지막으로 고객간 유사도를 기반으로 네트워크를 형성할 때 유사도 임계치 ρ(제 3.2절 참조)을 변화시킴으로써 서로 다른 표본을 추가로 획득했다. 이때 유사도 임계치 ρ은 0.1, 0.2만을 적용하였는데, 이는 ρ가 0.3이상에서는 네트워크의 연결이 희박해져 네트워크 구조를 분석하는 것이 의미가 없을 수 있기 때문이다.

이러한 과정을 통해 총 396(66×3×2)개의 최종 표본이 도출되었는데, 모형 구축을 위한 학습용 데이터(Training data)로 40%에 해당하는 표본을, 검증용 데이터(Test data)로 40%의 표본을, 그리고 나머지 20%를 평가용 데이터(Evaluation data)로 사용하였다. 또한 실험의 신뢰성을 높이기 위해 5-fold cross validation을 실시하였고, 이때 총 5개의 데이터셋(Fold)을 실험마다 다르게 추출하여 중복되지 않게 적용하였다.

3.2 독립변수 측정

독립변수로 사용되는 사회네트워크의 구조적 지표를 측정하는 프로세스는 <그림 1>과 같이 고객 유사도 분석, 사회 네트워크 구성, 사회 네트워크 구조적 지표 측정 등 총 3단계의 과정으로 이루어진다.



<그림 1> 독립변수 측정 프로세스

(1) 고객 유사도 분석

이 단계에서는 고객들의 구매패턴을 분석하고 실제 각 고객간의 구매패턴의 유사성을 측정한다. 구매패턴을 분석하기 위해서 구매 트랜잭션으로부터 고객이 어떠한 상품을 구매했는지를 파악하고 그 결과를 식 (1)과 같은 고객-상품 매트릭스 **P** = (p_{ij})로 표현한다. 단, 고객이 같은 상품을 여러 번 구매해도 1회 구매로 간주한다.

$$p_{ij} = \begin{cases} 1 : \text{고객 } i\text{가 상품 } j\text{를 구매} \\ 0 : \text{고객 } i\text{가 상품 } j\text{를 비구매} \end{cases} \quad (1)$$

고객-상품 매트릭스가 완성되면 식 (2)와 같이 *Jaccard* 유사도를 이용하여 고객들 사이의 유사성을 계산한다. 식 (2)에서 M_{11}은 고객 a와 b 모두 1의 값을 갖는 경우, M_{01}은 고객 a는 0, b는 1의 값을 갖는 경우, M_{10}은 고객 a는 1, b는 0의 값을 갖는 경우, M_{00}은 고객 a와 b 모두 0의 값을 갖는 경우를 나타낸다. *Jaccard* 유사도는 0과 1사의 값을 갖는데 1이면 두 고객 사이의 구매패턴이 완전히 일치한다는 것을 나타낸다.

$$sim(a, b) = J(a, b) = \frac{M_{11}}{M_{10} + M_{01} + M_{11}} \quad (2)$$

(2) 사회 네트워크 구성

앞 단계에서 계산한 유사도로 고객간 사회네트워크를 구성하는 단계이다. 고객간 유사도를 계산한 후 제품 구매 패턴이 유사한 고객들 사이의 네트워크를 구축한다. 네트워크는 동일한 제품을 구매한 고객들 사이에 관계가 있다고 가정하면 거의 모든 고객이 링크로 연결되는 문제점이 있기 때문에 고객간 유사도가 특정 임계치 ρ이상인 값을 1로 정의하여 고객들을 링크로 연결하였다. 고객 a와 b간의 연결여부 *link(a, b)*는 식 (3)과 같이 정의된다.

$$link(a, b) = \begin{cases} 1, sim(a, b) \geq \rho \\ 0, sim(a, b) < \rho \end{cases} \quad (3)$$

이와 같이 형성된 사회 네트워크에서는 비슷한 구매패턴을 가진 이웃이 링크로 연결된다. 앞에서 기술한 것처럼 실험에서는 임계치 ρ을 0.1과 0.2로 설정하여 사회 네트워크를 구성하였다. 0.3이상의 임계치에서는 고객간의 관계 자체가 너무 희박하기 때문이다.

(3) 사회 네트워크 구조적 지표 측정

앞 단계에서 구축된 사회 네트워크로부터 최종적으로 독립변수를 측정한다. 모형 구축에 사용할 독립변수는 제 2장에서 기술한 구조적 측정지표들로서 밀도, 포괄성, 집중도, 군집화계수, 효율성 등의 5가지이다. 이러한 지표들은 기존 연구에 근거한 수식을 통하여 계산되는데, 사회네트워크 분석 도구로 널리 사용되고 있는 UCINET 6.0과 NetMiner 3를 통하여 실제 값을 측정하였다.

포괄성(Inclusiveness)은 한 네트워크에 포함된 참여자의 총 수에서 연결되지 않은 참여자들의 수를 뺀 연결된 참여자들의 비율로 정의된다(Wasserman and Faust, 1994; 손동원, 2002). **밀도(Density)**는 한 네트워크의 참여자 간의 가능한 모든

관계 중에서 실제로 맺어진 관계 수의 비율로 정의된다(손동원, 2002; 김용학, 2003). 포괄성은 식 (4)와 같이, 밀도는 식 (5)와 같이 계산된다. 식 (5)에서 n은 네트워크 전체 노드의 수, k는 실제 연결된 관계의 수를 나타낸다.

$$\text{포괄성비율} = \frac{(\text{연결된 정의 수})}{(\text{네트워크 전체 점의 수})} \quad (4)$$

$$\text{밀도} = \frac{k}{n(n-1)/2} \quad (5)$$

집중도(Centralization)는 일반적으로 연결정도 집중도, 매개 집중도, 근접 집중도가 있다. 네트워크 내에서 중심적인 위치에 존재하는 정도를 나타내기 위해 연결정도 중심성, 근접 중심성, 매개 중심성를 통해 나타낸다. 연결정도 중심성은 한 참여자에 직접 관계를 맺는 노드들의 수를 의미한다. 해당 중심성이 높을수록 직접적인 관계의 참여자가 많다는 것을 의미하고 이를 근거로 네트워크 내에서 해당 참여자의 중심성 정도를 나타낸다. 근접 중심성은 한 참여자를 중심으로 직접적이거나 간접적으로 도달할 수 있는 모든 다른 참여자들과의 거리를 합산하여 나타낸다. 가장 짧은 거리를 가질수록 쉽게 다른 참여자에 영향을 줄 수 있다고 보고 중심적인 참여자로 본다. 매개 중심성은 네트워크 내의 다른 참여자들이 해당 참여자를 통해서 다른 참여자들과 연결되는 정도를 나타낸다. 이러한 매개 중심성이 높으면 다른 참여자들 사이의 경로에 많이 참여해 있다는 것을 나타내므로 네트워크에서 관계의 흐름을 위한 중요한 역할을 한다고 볼 수 있다(Freeman, 1979; Bonachich, 1987; 손동원, 2002). 집중도는 이러한 3가지의 중심성 지표의 관점에서 각각의 중심성이 높은 참여자에 얼마나 네트워크 관계가 집중되어 있는지를 나타낸다. 해당 값은 0에서 1사이의 값을 가지는데 네트워크의 구조가 높은 참여자에게 집중될수록 값은 1을 나타낸다. 네트워크 전체 참여자의 수를 n이라고 할 때, 연결정도 집중도는 식 (6)과 같은 방법을 통해 계산될 수 있다. $C_D(P^*)$은 네트워크에서 나올 수 있는 가장 높은 연결정도 중심성 값을 의미하는데 분자에서는 현재 네트워크에서 가장 높은 연결정도 중심성을 말하고 분모에서는 이론적으로 가능할 수 있는 최대치를 말한다. 따라서 분모의 값은 스타형 네트워크일 경우에 최대이므로 $(n-1)(n-2)$로 나타낼 수 있다. $C_D(P_i)$는 참여자 i의 연결정도 중심성 값을 말한다(Freeman, 1979; 손동원, 2002).

$$NC_D = \frac{\sum_{i=1}^{n}[C_D(P^*) - C_D(P_i)]}{max\sum_{i=1}^{n}[C_D(P^*) - C_D(P_i)]} = \frac{\sum_{i=1}^{n}[C_D(P^*) - C_D(P_i)]}{(n-1)(n-2)} \quad (6)$$

근접 집중도는 식 (7)과 같이 계산된다. 분모는 이론적으로 가능한 최대의 근접 집중도를 나타내며, 분자 중에서 $C_C(P_i)$는 참여자 i의 근접 중심성 값을 말한다(Freeman, 1979; 손동원, 2002).

$$NC_C = \frac{\sum_{i=1}^{n}[C_C(P^*) - C_C(P_i)]}{[(n-1)(n-2)/(2n-3)]} \quad (7)$$

매개 집중도는 식 (8)과 같은 방법으로 구할 수 있다. 분모는 네트워크에서 이론적으로 가능한 최대의 매개 집중도를 의미하며, 분자 중에서 $C_B(P_i)$는 참여자 i의 매개 중심성 값을 말한다. 편의상 식 (9)와 같이 분모를 표준화하여 계산한다(Freeman, 1979; 손동원, 2002).

$$NC_B = \frac{\sum_{i=1}^{n}[C_B(P^*) - C_B(P_i)]}{[(n-1)^2/(2n-3)]} \quad (8)$$

$$NC_B = \frac{\sum_{i=1}^{n}[C_B(P^*) - C_B(P_i)]}{[(n-1)]} \quad (9)$$

본 연구에서는 모형 구축과 분석의 편의를 위하여 각 집중도를 개별 변수로 사용하는 대신, 연결정도 집중도, 매개 집중도, 근접 집중도의 값을 평균하여 사용한다.

군집화계수(Clustering coefficient)는 네트워크 내의 3명의 참여자 *a, b, c*가 있고 *a*와 *b, a*와 *c* 사이에 관계가 있을 때 *b*와 *c*가 관계를 가질 가망성을 말한다. 예를 들자면 네트워크가 친구관계로 연결되어 있을 경우, 한 사람의 친구들이 서로 친구가 될 가능성으로 말할 수 있다. 여기서 *a*와 *b, a*와 *c* 사이에 관계가 있을 때를 삼자관계(triple)이라고 하며 *a, b, c* 사이에 모든 관계가 연결되어 있을 때를 삼각관계(triangle)라고 한다(Watts, 1999; Schank and Wagner, 2005). 군집화계수는 식 (10)과 같이 네트워크의 각 노드의 입장에서 존재하는 삼각관계의 수를 삼자관계의 수로 나누어 계산한다. 이러한 각 참여자 별 군집화계수의 값을 평균한 것이 네트워크 수준의 군집화계수이다. 전체 네트워크의 군집화계수는 식 (11)을 통하여 도출 할 수 있다(Schank and Wagner, 2005).

$$c(a) = \frac{a\text{의 삼각관계의 수}}{a\text{의 삼자관계의 수}} \quad (10)$$

$$CC = \frac{1}{|V|}\sum_{a \in V} c(a), \quad (11)$$

단 V는 연결정도가 2이상인 노드의 집합

효율성(Efficiency)은 네트워크에 존재하는 컴포넌트(각 참여자 *n*명이 끊기지 않고 연결된 집단)들에서 *n-1*의 연결이 필수적이고 가장 효율적인 연결의 수임을 의미한다(Krackhardt, 1994). 이는 식 (12)를 통해서 계산할 수 있는데, *V*는 이러한 모든 컴포넌트에서 효율적인 연결을 초과하는 연결수를 합한 것을 말하며, *Max(V)*는 모든 컴포넌트에서 가능한 최대 초과연결 수를 합산한 것을 나타낸다.

$$\text{효율성} = 1 - \frac{V}{Max(v)} \quad (12)$$

3.3 종속변수 측정

협업필터링 추천성과를 측정하기 위하여 대표적 협업필터링 알고리즘 중의 하나인 User-based Neighborhood 알고리즘(Sarwar et al., 2000; Huang et al., 2007)을 적용하여 추천시스템을 구축하였다. 협업필터링의 실제 성과를 측정하기 위하여 396개의 표본 각각을 추천 학습용 데이터(40일 동안의 구매 트랜잭션)와 추천 검증용 데이터(20일 동안의 구매 트랜잭션)로 분할한 후, 추천 학습용 데이터를 통해 고객들에게 추천할 상품을 정하고 추천 검증용 데이터에서의 실제 구매상품과 추천상품을 비교하여 추천 정확도를 측정하였다. 추천 정확도를 측정하기 위해 자주 사용되는 지표로는 재현율(*recall*), 정확율(*precision*), *F1-measure*가 있다. 재현율과 정확율은 추천 집합이 커질 때 재현율은 올라가고 정확율은 떨어지는 상반된 결과를 보이기 때문에, 본 연구에서는 식 (13)과 같이 두 지표를 동일한 가중치로 결합한 *F1-measure*를 사용하였다(Sarwar et al., 2000; 박종학 외, 2009).

$$F1 = \frac{recall \times precision}{recall + precision/2} \quad (13)$$

3.4 인공신경망 모형구축

인공신경망의 성능, 즉 최적화에 영향을 미치는

요인으로서는 은닉층의 수, 은익층의 노드 수, 학습반복횟수, 학습율, 모멘텀 등이 있는 것으로 알려져 있다(안현철 외, 2006; 박상길 외, 2006). 따라서 본 연구에서는 가장 널리 활용되고 있는 은익층이 하나인 3층 구조의 네트워크 모형을 사용하였고, 은익층의 노드의 수는 독립변수와 종속변수 개수의 합을 n이라 할 때 $n/2$, n, $3n/2$, $2n$의 총 4가지 경우인 3, 6, 9, 12로 구분하여 실험하였다. 편의상 학습율과 모멘텀은 0.1로 고정시켰고, 학습반복 횟수는 학습용 데이터 개수의 50배를 적용하여 16000회가 반복되면 학습을 중단하도록 하였다.

인공신경망 모형은 Clementine 11.1을 사용하여 분석하였는데, 신경망 생성방법을 Quick 모드로 하여 위의 모든 경우를 학습시킨 후 모형을 평가하여 최적 모형을 선택하였다. 최적 모형을 통한 예측 값과 실제 값 간의 오차 정도를 알아보기 위해 일반적으로 많이 사용되는 RMSE(Root Mean Squared Error)를 사용하였다. RMSE는 추정 모형이 실제 값을 추정하는데 얼마나 오차를 발생시키는 가를 평균적으로 나타내는 것으로 0에 가까울수록 모형이 실제 값을 잘 설명한다고 평가할 수 있다(Su and Khoshgoftaar, 2009).

4. 연구 결과

<표 1>과 같이 인공신경망 모형에 의한 협업필터링 추천의 예측력(설명력)는 5-fold cross validation을 위해 분류된 각 데이터셋에 대해 약 92.24%~93.27%(평균 92.61%) 정도로 매우 높게 나타났다. 또한 예측 값과 실제 값 간의 오차 정도인 RMSE는 약 0.00465~0.00584(평균 0.0049) 사이로 분석되었다. 이러한 결과로 미루어 볼 때 구축된 인공신경망 모형이 설명력과 정확도 측면 모두에서 만족할 만한 수준이라고 여겨진다. 따라서 본 연구에서 제시한 협업필터링 추천성능 예측모형은 기업에서 추천시스템 도입 여부를 판단하고자 할 때 유용하게 사용될 수 있을 것으로 판단된다. 즉, 추천성능 예측모형을 기업의 구매 데이터에 적용하여 적정한 수준의 RMSE가 나오면 협업필터링 추천시스템을 구축하고, 그렇지 않으면 다른 방식의 추천시스템을 대안으로 고려하거나 추천시스템 도입 자체를 폐기하는 것이 바람직할 것이다.

<표 1> 모형의 예측 정도

데이터	최적노드의 수	예측력	RMSE
데이터셋#1	6	92.24%	0.00476
데이터셋#2	3	92.29%	0.00484
데이터셋#3	3	93.27%	0.00548
데이터셋#4	9	92.54%	0.00496
데이터셋#5	3	92.69%	0.00465
전체 평균		92.61%	0.00490

<표 2> 입력 변수의 영향력

RMSE	입력변수의 상대적 중요도	
0.00465	밀도	0.385
	포괄성	0.445
	효율성	0.183
	집중도	0.159
	군집화계수	0.019

<표 2>는 RMSE가 가장 낮은 인공신경망 모형에서 각 독립변수가 종속변수에 미치는 상대적 영향력을 보여준다. 독립변수의 상대적 비교 정도를 보면 포괄성과 밀도가 다른 변수들 보다 월등히 높은 영향력을 미치는 것으로 분석되었다. 즉, 포괄성과 밀도 같은 네트워크의 구조적인 결속을 나타내는 지표들이 협업필터링 추천 성능을 결정짓는 가장 중요한 요인임을 알 수 있다. 이러한 결과는 구매 데이터로부터 파생된 사회네트워크에서

밀도나 포괄성 값이 낮은 경우에 협업필터링이 적절한 추천시스템이 될 수 없다는 것을 암시한다. 따라서 구매 데이터의 특성이 이와 같은 경우 이에 적합한 기존의 추천기법이 무엇인지를 탐색하거나 새로운 추천기법을 개발하는 연구가 추후에 진행될 필요가 있다.

5. 결 론

협업필터링 추천시스템을 구축하려면 상당한 시간과 비용이 소요된다. 따라서 추천시스템을 도입하고자 할 경우 그 성능이 어느 정도인지를 예측하는 일은 경제적 측면과 고객 만족도 측면에서 매우 중요하다. 본 연구에서는 협업필터링 추천시스템을 구축하기 않아도 추천시스템의 성능을 사전에 손쉽게 예측할 수 있는 인공신경망 모형을 제시하고, 실제 기업의 거래 데이터를 이용하여 모형을 구축하고 검증하였다. 인공신경망 모형의 입력변수(독립변수)를 추출하기 위해서 사회 네트워크 분석이 활용하였다. 실험 결과로부터 제시한 인공신경망 모형은 우수한 예측력을 보임으로써 협업필터링 추천 성능을 예측할 수 있는 기법으로 기업에서 추천시스템 도입 여부를 결정하고자 할 때 유용하게 사용될 수 있을 것으로 기대된다.

참고문헌

김용학, 사회연결망 분석, 박영사, 2003.

김재경, 조윤호, 김승태, 김혜경, “모바일 전자상거래 환경에 적합한 개인화된 추천시스템”, *경영정보학연구*, 15권 3호(2005), 223~241.

손동원, 사회네트워크 분석, 경문사, 2002.

박상길, 이해진, 심현진, 이정윤, 오재응, “회귀모형과 신경망모형을 이용한 차량공조시스템의 음질 인덱스 구축”, 한국소음진동공학회 춘계학술대회논문집, (2006), 1~6.

박종학, 조윤호, 김재경, “사회연결망 : 신규고객 추천문제의 새로운 접근법”, *지능정보연구*, 15권 1호(2009), 123~139.

안현철, 김경재, 한인구, “다분류 Support Vector Machine을 이용한 한국 기업의 지능형 기업채권평가모형”, 경영학연구, 35권 5호(2006), 1479~1496.

조영빈, 조윤호, “구매순서를 고려한 개선된 협업필터링 방법론”, 지능정보연구, 13권 2호(2007), 69~80.

조윤호, 방정혜, “신상품 추천을 위한 사회연결망 분석의 활용”, 지능정보연구, 15권 4호(2009), 183~200.

Adomavicious, G. and A. Tuzhilin, “Toward the next generation of recommender systems : A survey of the state-of-the-art and possible extensions”, *IEEE Transactions on Knowledge and Data Engineering*, Vol.17, No.6 (2005), 734~749.

Amorim, S., J. P. Barthelemy, and C. Ribeiro, “Clustering and clique partitioning : simulated annealing and tabu search approaches”, *Journal of Classification*, Vol.9(1992), 17~41.

Bonacich, P., “Power and Centrality : A Family of Measures”, *American Journal of Sociology*, Vol.92(1987), 1170~1182.

Breiger, R., S. Boorman, and P. Arabie, “An algorithm for clustering relational data, with applications to social network analysis and comparison with multi-dimensional scaling”, *Journal of Mathematical Psychology*, Vol.12(1975), 328~383.

Bron, C. and Kerbosch, J., ”Finding all cliques of an undirected graph”, *Communication of the ACM*, Vol.16(1973), 575~577.

Burt, R. S., Structure 4.1 Reference Manual, NY : Columbia University, 1991.

Burt, R. S., Structural Holes : The Social Structure of Competition, Cambridge, MA : Harvard University Press, 1992.

Cho, Y. H. and J. K. Kim, "Application of web usage mining and product taxonomy to collaborative recommendations in e-commerce", Expert Systems with Applications, Vol.26, No.3(2004), 234~246.

Frank, O. and F. Harary, "Cluster Inference by Using Transitivity Indices in Empirical Graphs", *Journal of the American Statistical Association*, Vol.77, No.380(1982), 835~840.

Freeman, L., "Centrality in Social Networks : Conceptual clarification", *Social Networks*, Vol.1(1979), 215~239.

Huang, Z. and D. Zeng, "Why Does Collaborative Filtering Work?-Recommendation Model Validation and Selection by Analyzing Random Bipartite Graphs", the Fifteenth Annual Workshop on Information Technologies and Systems, 2005.

Huang, Z., D. Zeng, and H. Chen, "A Comparative Study of Recommendation Algorithms in E-commerce Applications", *IEEE Intelligent Systems*, Vol.22, No.5(2007), 68~78.

Human, S. E. and K. G. Provan, "Legitimacy Building in the Evolution of Small-Firm Multilateral Networks : A Comparative Study of Success and Demise", *Administrative Science Quarterly*, Vol.45, No.2(2000), 327~365.

Krackhardt, D., "Graph Theorical Dimentions of Informal Organizations", In Kathleen Carley and Michael Prietula(eds.), Computational Organizational Theory, Hillsdale, NJ : Lawrence Erilbaum Associations, (1994), 89~111.

Kukkonen, H. O., K. Lyytinen, and Y. J. Yoo, "Social Networks and Information Systems : Ongoing and Future Research Streams", *Journal of the Association for Information Systems*, Vol.11(2010), 61~68.

Liu, F. and H. J. Lee, "Use of social network information to enhance collaborative filtering performance", *Expert Systems with Applications*, Vol.37(2010), 4772~4778.

Melville, P., R. J. Mooney, and R. Nagarajan, "Content-boosted Collaborative Filtering", Proceeding SIGIR 2001 Workshop on Recommender Systems, 2001.

Oh, W., J. Choi, and K. Kim, "Coauthorship Dynamics and Knowledge Capital : The Patterns of Cross-disciplinary Collaboration in Information System Research", *Journal of Management Information System*, Vol.22, No.3(2006), 265~292.

Ryu, Y. U., H. K. Kim, Y. H. Cho, and J. K. Kim, "Peer-oriented content recommendation in a social network", Proceedings of the Sixteenth Workshop on Information Technologies and Systems, (2006), 115~120.

Sarwar, B., G. Karypis, J. A. Konstan, and J. Riedl, "Analysis of recommendation algorithms for e-commerce", Proceedings of ACM E-commerce Conference, (2000), 158~167.

Schank, T. and D. Wagner, "Approximating Clustering Coefficient and Transitivity", *JGAA*, Vol.9, No.2(2005), 265~275.

Scott, J., Social Network Analysis : A Handbook, Thousand Oaks, CA : Sage, 2000.

Seidman, S. B. and B. L. Foster, "A note on the potential for genuine cross-fertilization between anthropology and mathematics", *Social Networks*, Vol.1(1978), 65~72.

Su, X. and T. M. Khoshgoftaar, "A Survey of Collaborative Filtering Techniques", *Advances in Artificial Intelligence*, Vol.2009(2009).

Wasserman, S. and K. Faust, "Social Network

Analysis : Methods and Application", New-york : Cambridge University Press, 1994.

Watts, D. J., Small worlds, Princeton University Press, Princeton, New Jersey, 1999.

Weare, C., Loges, W. E., Oztas, N., "Email Effects on the Structure of Local Associations : A Social Network Analysis", *Social Science Quarterly*, Vol.88, No.1(2007), 222~243.

Abstract

Predicting the Performance of Recommender Systems through Social Network Analysis and Artificial Neural Network

Yoonho Cho* · Inhwan Kim**

The recommender system is one of the possible solutions to assist customers in finding the items they would like to purchase. To date, a variety of recommendation techniques have been developed. One of the most successful recommendation techniques is Collaborative Filtering (CF) that has been used in a number of different applications such as recommending Web pages, movies, music, articles and products. CF identifies customers whose tastes are similar to those of a given customer, and recommends items those customers have liked in the past. Numerous CF algorithms have been developed to increase the performance of recommender systems. Broadly, there are memory-based CF algorithms, model-based CF algorithms, and hybrid CF algorithms which combine CF with content-based techniques or other recommender systems. While many researchers have focused their efforts in improving CF performance, the theoretical justification of CF algorithms is lacking. That is, we do not know many things about how CF is done. Furthermore, the relative performances of CF algorithms are known to be domain and data dependent. It is very time-consuming and expensive to implement and launce a CF recommender system, and also the system unsuited for the given domain provides customers with poor quality recommendations that make them easily annoyed. Therefore, predicting the performances of CF algorithms in advance is practically important and needed.

In this study, we propose an efficient approach to predict the performance of CF. Social Network Analysis (SNA) and Artificial Neural Network (ANN) are applied to develop our prediction model. CF can be modeled as a social network in which customers are nodes and purchase relationships between customers are links. SNA facilitates an exploration of the topological properties of the network structure that are implicit in data for CF recommendations. An ANN model is developed through an analysis of network topology, such as network density, inclusiveness, clustering coefficient, network centralization, and Krackhardt's efficiency. While network density, expressed as

* School of Management Information Systems, Kookmin University
** Department of e-Business, Graduate School of Kookmin University

a proportion of the maximum possible number of links, captures the density of the whole network, the clustering coefficient captures the degree to which the overall network contains localized pockets of dense connectivity. Inclusiveness refers to the number of nodes which are included within the various connected parts of the social network. Centralization reflects the extent to which connections are concentrated in a small number of nodes rather than distributed equally among all nodes. Krackhardt's efficiency characterizes how dense the social network is beyond that barely needed to keep the social group even indirectly connected to one another. We use these social network measures as input variables of the ANN model. As an output variable, we use the recommendation accuracy measured by F1-measure.

In order to evaluate the effectiveness of the ANN model, sales transaction data from H department store, one of the well-known department stores in Korea, was used. Total 396 experimental samples were gathered, and we used 40%, 40%, and 20% of them, for training, test, and validation, respectively. The 5-fold cross validation was also conducted to enhance the reliability of our experiments. The input variable measuring process consists of following three steps; analysis of customer similarities, construction of a social network, and analysis of social network patterns. We used Net Miner 3 and UCINET 6.0 for SNA, and Clementine 11.1 for ANN modeling. The experiments reported that the ANN model has 92.61% estimated accuracy and 0.0049 RMSE. Thus, we can know that our prediction model helps decide whether CF is useful for a given application with certain data characteristics.

Key Words : Social Network Analysis, Collaborative Filtering, Neural Network, Recommendation Performance Prediction

저 자 소 개

조윤호

서울대학교 계산통계학과(전산학전공)를 졸업하고, KAIST 경영정보공학과에서 석사, KAIST 경영공학과에서 박사학위를 취득하였으며, LG전자(주)에서 6년간 주임연구원으로 재직하였다. 현재 국민대학교 경영대학 경영정보학부 부교수로 재직 중이다. 주 연구분야는 추천시스템, 모바일비즈니스, 고객관계관리, 데이터마이닝 등이다.

김인환

국민대학교 경영학부 e-비즈니스전공으로 학사학위를 받았으며 현재 국민대학교 대학원 e-비즈니스학과 석사과정에 재학 중이다. 주 연구분야는 social network analysis, 추천시스템, 데이터마이닝 등이며, 한국경영정보학회와 한국지능정보시스템학회 학술대회에서 추천시스템 분야의 논문을 발표하였다.