

사회연결망 분석을 활용한 상품추천방법의 성능향상 방안

강부식*

요약

급증하는 전자상거래 환경에서 상품추천시스템의 추천 성능 향상은 주요 이슈 중 하나이다. 협업 필터링 방식은 추천시스템에서 가장 성공적인 방식으로 일반적으로 많이 활용되고 있다. 협업 필터링 방식은 근본적으로 신규고객 및 신규제품 추천이 어려운 문제점이 있으며, 최근 사회연결망 분석을 활용하여 이를 해결하는 연구가 제시되고 있다. 사회연결망에는 다양한 분석기법이 있는 데, 연결정도 중심성과 구조적 공백 기법이 상품추천 분야에서 주로 적용되고 있다. 이 연구에서는 연결정도 중심성과 구조적 공백 기법을 협업 필터링 기반 추천 프로세스에 활용함으로써 추천 성능을 높이는 방안을 제시하였다. 영화 추천 데이터에 대한 실험결과를 통해 구조적 공백 기법을 협업 필터링 기반 추천 프로세스에 결합하여 활용함으로써 추천 성능을 높일 수 있음을 보였다.

Performance Improvement of Product Recommendation Methodology Using Social Network Analysis

Boo-Sik Kang*

ABSTRACT

In the E-commerce environment, performance improvement is one of the main issues in the field of product recommendation systems. Collaborative filtering is a successful and popular method for recommender systems. It has some fundamental problems for new customer or new product recommendation. Recently, some studies using social network analysis has been presented to solve the problems. There are various analysis methods in the field of the social networks. Degree centrality and structural holes of them has been used mainly for recommender systems. This study suggested a methodology for performance improvement of the collaborative filtering process using degree centrality and structural holes methods. In experimental results for MovieLens datasets, this study showed the proposed methodology using structural holes can improve the performance of the recommendation system.

Key Words : Recommendation System, Collaborative Filtering, Social Networks, Structural Holes, Degree Centrality, Performance Improvement

* 목원대학교 서비스경영학부(☐bookang@mokwon.ac.kr)

· 제1저자(First Author) : 강부식 · 교신저자(Correspondent Author) : 강부식

· 접수일(2011년 3월 8일), 수정일(1차 : 2011년 4월 8일), 게재확정일(2011년 4월 13일)

1. 서론

상품추천시스템은 통계적 기법과 데이터 마이닝 기법을 이용하여 고객 개개인의 취향에 가장 적합한 상품을 추천해 주는 시스템으로, 급증하는 전자상거래 환경에서 고객의 구매편의를 도모하고 매출증대에 기여할 수 있다[1,2]. 상품추천시스템은 연관규칙이나 군집분석[3], 사례기반추론[4] 등 다양한 방법에 의해 구현될 수 있는데, 협업 필터링 방식이 가장 성공적인 기법으로 알려져 있으며, 일반적으로 가장 많이 활용되고 있다[1,2,5,6]. 협업 필터링 방식은 추천 대상고객과 가장 유사한 성향을 보이는 이웃고객들을 찾아내고, 이웃고객들의 구매정보를 활용하여 가장 적중률이 높을 것으로 예상되는 추천 상품목록을 추출하여 이 상품들을 추천 대상고객에게 추천한다[1,6].

협업 필터링 방식은 상품추천을 위한 가장 성공적인 기법임에도 불구하고 신규고객 추천이나 신상품 추천을 할 수 없는 등의 몇 가지 근본적인 문제점을 가지고 있다. 신규고객의 경우에는 시스템 내에 구매성향에 대한 정보가 없기 때문에 이웃고객을 선정할 수 없으므로 협업 필터링 방식을 적용할 수 없고, 신제품의 경우에는 이웃고객들 중 아직까지 아무도 구매하지 않았으므로 추천 상품목록에 선정되지 못함으로 인해 협업 필터링 방식을 적용할 수 없다. 최근 사회연결망 기법을 활용하여 협업 필터링이 가지는 신규고객 문제[5,7]와 신제품 문제[2]를 해결하기 위한 연구가 제시되고 있다. 사회연결망(social network)은 사람들이 연결되어 있는 관계망으로, 사회연결망 분석으로 구조나 연결망 형태의 특징을 도출하고, 관계성으로 체계의 특성을 설명하거나 체계를 구성하는 단위의 행위를 설명할 수 있다[8,9]. 협업 필터링 방식도 고객의 구매 정보를 분석하여 준연결망을 생성한 후에 링크로 연결된 이웃고객들의 구매정보를 이용하여 특정 고객에게 상품을 추천할 수 있게 만든 일종의 사회연결망이라 할 수 있다[5].

상품추천시스템의 추천 성능은 전자상거래 환경하의 고객만족과 매출증대에 직접적으로 연계될 수 있는 주요 이슈이다[1,10]. 이 연구에서는 추천 성능을 높이기 위해 협업 필터링 방식에 사회연결망 분석 기법을 활용하는 방안에 대해 살펴본다. 박중학 등[5]은 신규고객에게 상품추천 시에 사회연결망의 연결정도 중심성이 성과가 좋았음을 제시하였다. 조윤희와 방정혜[2]는 신상품추천을 위한 사회연결망 연구에서 연결정도 중심성을 활용할 경우 비교적 높은 적중률을 보임을 실험결과를 통해 검증하였다. 강부식[7]은 신규고객을 위한 상품추천에서 사회연결망의 구조적 공백을 활용할 경우 베스트셀러 방식보다 추천 성능이 높음을 보였다. 이와 같은 사회연결망의 활용 연구는 기존 고객과 기존 상품추천을 위한 협업 필터링 방식의 추천 성능 향상에도 기여할 것으로 판단된다. 이에 이 연구에서는 협업 필터링 추천 프로세스에 사회연결망의 연결정도 중심성과 구조적 공백 기법을 활용하는 방안을 모색하고, 영화 추천 데이터를 통해 실험함으로써 추천 성능을 살펴보고자 한다.

II. 사회연결망 분석과 협업 필터링

2.1 사회연결망 분석

사회연결망 분석을 위해서는 행위자들의 관계에 대한 데이터가 필요하며, 관계를 이용하여 네트워크를 구성한다. 네트워크는 종종 매트릭스로 표현되며, 행위자간 관계가 있으면 1, 없으면 0으로 나타낼 수 있고, 혹은 각 관계의 정도를 수치로 표현할 수도 있다[8].

n 명의 사람들 사이의 관계망을 분석하기 위해서는 $(n \times n)$ 형태의 전체 네트워크를 행렬로 구성하고 이들 간의 관계를 행렬안의 각 셀에 표현하여 관계성을 분석한다. 이때 행과 열은 같은 개체(사람)로 구성되는데, 이를 1원(1-mode) 자료라고 부른다. 이 때 셀 c_{ij} 는 사람 i 와 j 간의 관계성을 나타낸다. 그러나 행과

열에 다른 수준의 개체가 배열될 경우가 있는데, 이를 2원(2-mode) 자료라고 부른다. 가령 i 라는 사람이 k 라는 책을 구입하면 $b_{ik}=1$ 로, 구입하지 않았으면 $b_{ik}=0$ 으로 표현한 행렬 B는 (사람*책)의 2원 자료가 된다. 행렬 B로부터 책을 매개로 사람사이의 관계를 도출하거나 사람을 매개로 책 사이의 관계를 도출할 수 있는데, 이렇게 직접적인 상호작용의 관계가 없더라도 관계를 인위적으로 설정한 연결망을 준연결망(quasi-network)이라 한다. 준연결망도 연결망의 대상이 되며, 사회과학에서는 기존의 자료를 네트워크 자료로 재구성하여 활용할 수 있다는 장점 때문에 널리 쓰인다[9]. 구매데이터로부터 고객연결망과 상품연결망을 구성하는 사례를 그림 1에서 살펴볼 수 있다[5].

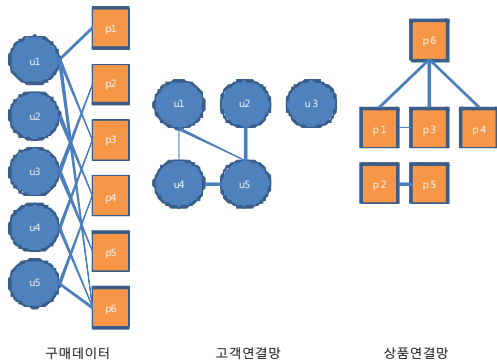


그림 1. 고객연결망과 상품연결망 구성
 Fig. 1. Construction of Customer Social Network and Product Social Network

사회연결망분석에서 관계 및 구조적 특성을 분석하기 위해 널리 사용되고 있는 기법으로 중심성(centrality), 밀도(density), 구조적 공백(혹은 구조적 틈새: structural hole), 중심화(centralization) 등이 있다[8,9]. 밀도는 가능한 총 관계의 수에서 실제로 맺어진 관계 수의 비율을 의미하며, 중심화는 전체 연결망의 형태가 어느 정도 중앙에 집중되었는지를 나타내는 개념이다[9]. 밀도와 중심화는 주로 연결망 전체를

대상으로 하는 척도이며, 연결망 내 각 노드에 대한 분석을 위해서는 중심성과 구조적 공백을 활용하여 분석할 수 있다[2,5,7].

중심성은 사회연결망 내에서 한 주체가 중심에 위치하는 정도를 나타내는 척도로 연결정도 중심성(degree centrality), 근접 중심성(closeness centrality), 사이 중심성(betweenness centrality), 위세 중심성(prestance centrality)이 있다[8,9]. 박종학 등[5]은 협력 필터링 방식은 추천대상고객과 직접 연결된 이웃고객과의 관계만 고려하고, 대상고객의 이웃의 이웃까지의 관계는 고려하지 않기 때문에 연결정도 중심성만을 적용하여 신규고객의 이웃고객을 찾는 방안을 제시하고 있다. 조윤희와 방정혜[2]의 신상품추천을 위한 사회연결망 활용 방안 연구에서는 위의 4가지 중심성 척도를 적용해 본 결과 연결정도 중심성이 가장 적중률이 높음을 보이고 있다. 연결정도 중심성은 한 노드에 직접적으로 연결되어 있는 노드들의 관계 값의 합으로 측정할 수 있는데, 이는 연결망에서의 각 노드의 중심성을 나타낸다[8,9].

중심성이 연결망 내에서 한 고객이 얼마나 많은 고객과 관계를 맺고 있는 가를 측정하는 척도라면 구조적 공백(structural holes)은 한 고객이 얼마나 다양한 고객들과 비중복적인 관계를 맺을 수 있는 위치에 있는가와 관련된 개념이다[8,9]. 예를 들어 사회연결망 내에서 사업가 5명을 만나는 것과 사업가, 정치인, 언론인, 학계, 체육계 각 1명씩 5명과 관계를 맺고 있는 것은 연결정도 중심성에서는 같지만 비중복적 관계를 맺고 있는 경우에 정보 확보 능력이 우월함으로 다양한 장점을 가질 수 있다는 것이다[8,9]. 구조적 공백 개념은 Burt[11]가 제시한 것으로, Burt는 비중복성과 낮은 제약성이라는 두 가지 개념으로 측정하려 했는데, 낮은 제약성이란 중복된 관계에 놓이지 않을 때 생긴다[9]. Burt는 구조적 공백위치를 확보한 소유자가 승진이 더 빠르고, 상사로부터 좋은 평가를 받으며, 더 좋은 보상을 받는다는 점을 밝혔다[9]. 강부식[7]은 신

규고객을 위한 상품추천 시에 구조적 공백 위치를 차지한 고객을 이웃고객으로 선정하여 상품추천 할 경우 추천 효과가 높음을 보였다.

2.2 협업 필터링

협업 필터링 방식은 추천 대상고객과 성향이 유사한 이웃고객들을 찾아내고, 이웃고객들의 구매정보를 활용하여 추천 대상고객이 구매하지 않은 상품들 중에서 가장 구매가능성이 높게 예측되는 상품들을 추천한다. 일반적으로 협업 필터링 기반 추천 프로세스는 고객유사도 계산, 이웃고객 선정, 상품별 구매가능성 예측점수 계산, 상품추천으로 구성된다[6, 12].

추천 대상고객의 이웃고객을 선정하기 위해, 먼저 구매데이터를 이용하여 고객 간 유사도를 측정하고 추천 대상고객과 가장 유사성이 높은 순으로 K명의 이웃을 선정한다. 협업 필터링 추천시스템은 고객의 구매데이터를 분석하여 준연결망을 생성한 후에 링크로 연결된 이웃고객들의 구매정보를 이용하여 특정 고객에게 상품을 추천할 수 있게 만든 일종의 사회연결망이라고 할 수 있다[5]. 협업 필터링은 연결망 내에서 추천 대상고객과 직접 연결된 고객들만을 대상으로 이웃을 선정하는 데, 연결망 내의 개체는 연결망 전체에서 영향력이 큰 개체들의 영향을 받을 수밖에 없다. 따라서 전체 연결망에서 영향력이 큰 고객들을 일부 이웃으로 선정하는 것은 이웃 선정의 좋은 방안이 될 수 있을 것으로 보인다.

상품추천을 위해서는 추천 대상고객이 구매하지 않은 상품들 중에서 이웃들의 구매 정보를 활용하여 구매가능성 예측점수를 계산하고, 구매가능성 점수가 높은 상품 순으로 N개를 추천한다. 일반적으로 협업 필터링에서 상품별 구매가능 예측점수는 고객의 구매 성향 정보만을 활용하여 계산하고 있고 상품 자체의 중요도 정도는 반영하지 않는다. 따라서 상품연결망으로부터 상품의 중요도를 찾아내고 이를 상품 구매 가능 예측 시 활용함으로써 추천 성능을 높일 수 있을

것으로 보인다.

고객연결망에서 영향력이 높은 고객이나 상품연결망에서 중요도가 높은 상품을 선정하기 위해 연결정도 중심성이나 구조적 공백 기법이 활용될 수 있다. 이 연구에서는 협업 필터링의 추천 프로세스에 이를 반영하는 방안을 통해 추천 성능을 높이는 방안을 제시하고자 한다.

III. 사회연결망을 활용한 추천 프로세스

협업 필터링 추천 프로세스에 사회연결망을 반영한 프로세스는 그림 2처럼 구성된다.

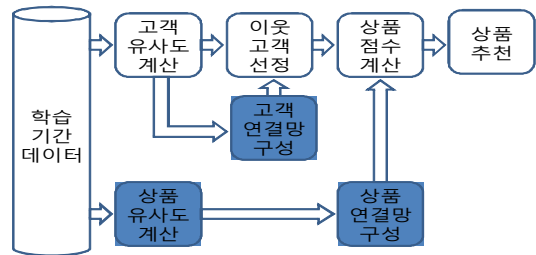


그림 2. 상품추천 프로세스
Fig. 2. Product Recommendation Process

3.1 고객유사도 계산

이 단계에서는 고객 구매데이터로부터 고객 간 유사도를 측정한다. 먼저, 고객 구매 데이터에서 특정 고객이 특정 상품을 구매한 경우 1, 구매하지 않은 경우 0으로 표현한다. 이를 고객-상품 행렬 $V = (v_{ij})$ 로 나타내면 식(1)과 같이 표현할 수 있다.

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

다음 고객-상품 매트릭스를 이용하여 고객 간 유사

도를 측정한다. 유사도를 측정하기 위해 많이 사용하는 기법은 피어슨 상관계수이다[1,2,5,6]. 식(1)과 같이 개체가 이진 값을 갖는 경우 유사도를 측정하는 또 하나의 방법은 조윤호와 김인환[13]의 연구에서처럼 식(2)와 같은 Jaccard 유사도를 이용하는 것이다. 식(2)에서 M_{11} 은 고객 a 와 b 가 모두 구매한 상품의 수, M_{10} 은 고객 a 만 구매한 상품 수, M_{01} 은 고객 b 만 구매한 상품 수를 나타낸다. Jaccard 유사도는 0과 1사이의 값을 가지며, 1은 두 고객의 특성이 일치함을 의미한다.

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

본 연구에서는 이진 값의 경우 유사도 계산이 간단한 Jaccard 유사도 척도를 사용하여 상품 간 유사도를 측정한다.

3.2 이웃고객 선정

추천 대상고객을 위한 이웃고객 K 명을 선정하는 단계이다. 협업 필터링 방식은 추천 대상고객과 유사도가 큰 순서로 K 명의 고객을 선정한다. 이 연구에서 제안한 방안은 유사도가 큰 순서로 $(K/2)$ 명, 고객연결망 내의 영향력이 큰 순서로 $(K/2)$ 명을 3.2.1의 단계처럼 선정한다.

3.2.1 고객연결망 구성 및 중심고객 선정

(1) 고객연결망 구성

고객유사도를 이용하여 고객연결망을 구성한다. 고객연결망을 구성하는 한 가지 방법은 고객유사도가 특정 임계값 ρ 이상이면 고객 간을 연결하는 링크의 가중치를 식(3)과 같이 1로 정의하고, 링크의 가중치가 1인 경우에만 두 고객을 연결하여 고객연결망을 구성하는 것이다[2,5,7].

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

식(3)를 이용하여 고객연결망을 구성하는 경우 적절한 ρ 를 찾기 위해 반복된 많은 실험을 필요로 하고, 원래의 고객유사도가 가지고 있는 정보손실이 발생하는 문제점이 있다. 원래의 고객유사도를 이용하여 고객연결망을 구성하는 경우 이 문제점을 피할 수 있다. 따라서 본 연구에서는 식(4)를 이용하여 고객연결망을 구성한다.

$$link(a, b) = s(a, b) \quad (4)$$

(2) 중심고객 선정

중심고객은 고객연결망에서 영향력이 큰 고객을 의미한다. 중심고객을 선정하기 위해 연결정도 중심성과 구조적 공백 기법을 활용한다.

고객연결망에서 각 고객의 연결정도 중심성은 다른 고객과 연결된 링크 가중치의 합으로 측정된다. 고객 a 의 연결정도 중심성 $cen(a)$ 는 식(5)와 같이 정의된다. N 은 고객 a 에 연결된 고객들이 N 명임을 의미한다.

$$cen(a) = \sum_{j=1}^N link(a, j) \quad (5)$$

구조적 공백 개념은 Burt[11]가 제시한 것으로, Burt는 연결망에서의 유리한 위치를 비중복성과 낮은 제약성이라는 개념으로 측정하려고 했다[9]. 연결망내의 개체 i 와 j 의 제약성 c_{ij} 는 식(6)과 같이 계산한다[11].

$$c_{ij} = \left(p_{ij} + \sum_{k, k \neq i, k \neq j} p_{ik} p_{kj} \right)^2 \quad (6)$$

$$p_{ij} = \frac{a_{ij} + a_{ji}}{\sum_k (a_{ik} + a_{ki})} \quad (7)$$

p_{ij} 는 개체 i 가 연결망의 모든 개체와 갖는 관계의 총합에 대해 개체 i 가 개체 j 와 갖는 관계가 차지하는 비율을 나타낸다. a_{ij} 는 개체 i 로부터 개체 j 로 향하는 링크의 값을 나타낸다. 방향성이 없는 링크의 경우에는 $a_{ij} = a_{ji}$ 이다. 식(6)에서 제약성이 커지는 이유는 크게 2가지가 있다[9]. 개체 i 가 개체 j 에 투자하는 비율의 정도와 개체 i 가 제 3자 개체인 k 에게 투자를 많이 하는데, 3자 개체인 k 가 개체 j 에게 투자를 많이 할수록 제약성이 커지게 된다. 한 연결망에서 개체 i 의 제약성의 총합은 식(8)과 같이 구할 수 있다[11]. 만약 i 가 고립개체인 경우에는 $C_i = 1$ 이 된다.

$$C_i = \sum_j c_{ij} \quad (8)$$

낮은 제약성은 구조적 공백을 연결할 경우에 생기며, 따라서 개체 i 가 얼마나 구조적 공백의 위치를 잘 활용하고 있는지를 나타내면 식(9)와 같이 정의할 수 있다.

$$str(i) = 1/C_i \quad (9)$$

고객연결망에서 고객 a 의 영향력을 $Imp(a)$ 라 하면, $Imp(a)$ 는 식(10)과 같이 나타낼 수 있다.

$$Imp(a) = \begin{cases} cen(a), & \text{연결정도 중심성} \\ str(a), & \text{구조적 공백} \end{cases} \quad (10)$$

$Imp(a)$ 수치가 높을수록 연결망에서의 영향력이 높다고 판단할 수 있으며, 영향력이 높은 ($K/2$)명의 고객을 중심고객, 즉 이웃고객으로 선정한다.

3.3 상품점수 계산

이웃고객 K 명의 구매정보를 활용하여, 추천 대상 고객 a 가 구매하지 않은 상품 c 를 대상으로 상품 구매 가능 예측점수를 식(11)과 같이 계산한다.

$$pls(a, c) = \frac{\sum_{j \in K} v_{jc} \times w(a, j)}{\sum_{j \in K} w(a, j)} \quad (11)$$

v_{jc} 는 이웃고객 j 의 상품 c 에 대한 구매여부를 나타내며(구매한 경우 1, 구매하지 않은 경우 0), K 는 이웃고객 집합을, $w(a, j)$ 는 추천 대상고객 a 와 이웃고객 j 의 관계 가중치를 의미하며, 일반적으로 협업 필터링에서는 두 고객 사이의 유사도인 $s(a, j)$ 를 사용한다. pls 계산에 상품의 중요도를 반영하기 위해서는 상품연결망을 구성하고 상품별 중요도를 측정한다.

3.3.1 상품유사도 계산

3.1의 고객-상품 매트릭스를 이용하여 상품유사도를 식(12)와 같이 측정한다. 식(12)에서 M_{11} 은 상품 x 와 y 를 모두 구매한 고객의 수, M_{10} 은 상품 x 만 구매한 고객 수, M_{01} 은 상품 y 만 구매한 고객 수를 나타낸다.

$$s(x, y) = J(x, y) = \frac{M_{11}}{M_{10} + M_{01} + M_{11}} \quad (12)$$

3.3.2 상품연결망 구성 및 상품별 중요도 계산

상품연결망 구성은 식(4)와 유사하게 상품유사도가 상품 간 링크의 가중치가 되도록 구성한다(식(13) 참조).

$$link(x, y) = s(x, y) \quad (13)$$

상품연결망의 연결정도 중심성은 식(5)와 유사하게 식(14)처럼 정의된다.

$$cen(x) = \sum_{j=1}^N link(x, j) \quad (14)$$

구조적 공백은 식(9)를 이용해 구한다. 각 상품의 중요도 $Imp(x)$ 는 식(10)과 유사하게 식(15)를 이용해 계산한다.

$$Imp(x) = \begin{cases} cen(x), & \text{연결정도 중심성} \\ str(x), & \text{구조적 공백} \end{cases} \quad (15)$$

$Imp(x)$ 는 상품 x 의 중요도이며, 이를 이용하여 식(11)의 상품별 구매가능 예측점수 pls 를 식(16)과 같이 보정할 수 있다.

$$pls(a, c) = \left(\frac{\sum_{j \in K} v_{jc} \times w(a, j)}{\sum_{j \in K} w(a, j)} \right) \times Imp(c) \quad (16)$$

3.4 상품추천

추천 대상고객이 구매하지 않은 상품 중에서 구매 가능 예측점수가 높은 순으로 N 개의 상품을 추천한다.

IV. 실험설계

4.1 실험 데이터 및 실험 방법

본 연구에서 제안한 상품추천 방안에 대한 검증을 위해 영화추천 시스템 데이터인 MovieLens (<http://movielens.umn.edu>)를 실험하였다. 이 데이터는 미네소타 대학의 GroupLens Research Project팀에 의해 수집되었으며, 1997년 9월부터 1998년 4월까지 943명의 사용자가 1682편의 영화에

대해 평가한 100,000개의 데이터가 저장되어 있다. 사회연결망 분석은 대형 사회연결망 분석을 위해 만들어진 무료 프로그램인 Pajek[14]을 이용하였다. 실험 분석을 위해서 Visual Basic과 Access를 사용하였다. 추천 성능 평가를 위해 다음 3가지 방안에 대해 실험하였다.

가. 협업 필터링(CF)

나. 연결정도 중심성을 활용한 협업 필터링(CFD)

다. 구조적 공백을 활용한 협업 필터링(CFH)

구체적인 실험방법은 다음과 같다.

1) 100,000개의 레코드를 데이터가 입력된 시간 순으로 정렬한 다음, 50,000개의 학습기간 데이터와 50,000개의 적용기간 데이터로 분할하였다. 학습기간에는 491명의 사용자와 1,466개의 영화가 포함되었다.

2) 491명의 각 사용자가 본 영화는 1, 보지 않은 영화는 0의 값을 갖도록 $491 \times 1,466$ 사용자-영화 매트릭스를 구성하고(식(1) 참조), 식(2)를 이용하여 491명의 두 사용자 사이의 유사도를 측정하였다.

3) 사용자 유사도가 식(4)와 같이 사용자 사이를 연결하는 링크의 값이 되도록 고객연결망을 구성하였다.

4) 고객연결망에서 식(5)를 이용하여 개별 고객의 연결정도 중심성을 계산하고, 식(9)를 이용하여 구조적 공백을 계산하였다.

5) 사용자-영화 매트릭스에서 식(12)를 이용하여 영화 간 유사도를 계산하고, 식(13)를 이용하여 상품연결망을 구성한 다음 개별 상품의 연결정도 중심성과 구조적 공백을 측정하였다.

6) CF에서는 추천 대상고객 a 에 대해 a 와 유사도가 큰 K 명을 선정하였다. CFD에서는 유사도가 큰 $(K/2)$ 명과 연결정도 중심성이 큰 고객 $(K/2)$ 명을 이웃고객으로 선정하였다. CFH는 유사도가 큰 $(K/2)$ 명과 구조적 공백이 큰 고객 $(K/2)$ 명을 이웃고객으로 선정하였다. 이웃고객의 수 K 는 10에서

100까지 10씩 증가하며 실험하였다.

7) 추천 대상고객이 보지 않은 영화에 대해 *pls*를 계산하기 위해 CF는 식(11)을 사용하였다. 고객 간 관계 가중치인 $w(a,j)$ 는 고객유사도인 $s(a,j)$ 를 사용하였다. CFI는 식(16)을 사용하고 각 영화의 중요도는 영화의 연결정도 중심성 값을 이용하였다. CFH는 *pls* 계산을 위해 식(16)을, 영화의 중요도는 개별 영화의 구조적 공백 값을 이용하였다. CFD와 CFH의 경우 이웃고객을 선정한 경로가 다르므로 추천 대상고객 a 와 선정된 이웃고객 j 사이의 관계가중치인 $w(a,j)$ 는 1로 설정하여 사용하였다.

8) 추천 대상고객에게 *pls* 점수가 높은 순으로 N 개의 영화를 추천하였다. N 은 10, 15, 20 세 가지 경우에 대해 실험하였다.

9) 적용기간 데이터는 성능 평가를 위해 사용되었다. 적용기간 데이터 중에서 신규 영화 및 신규고객에 해당하는 데이터를 제거한 후의 5993개의 데이터(사용자 134명, 영화 1086개 포함)가 검증데이터로 사용되었다.

4.2 성능평가지표

추천 시스템의 추천정확성 평가를 위해 일반적으로 사용하는 척도는 F1으로, 본 연구에서도 F1을 사용하였다.

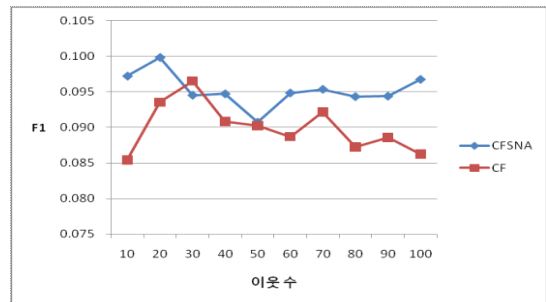
$$F1 = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (17)$$

Recall은 검증데이터내의 상품 집합 중에서 추천 상품 집합에 속하게 된 상품의 비율을 나타내고, Precision은 추천시스템에서 생성한 추천 상품 집합 중에서 검증데이터내의 상품 집합에 속하는 상품의 비율을 의미한다. 추천집합의 크기가 증가함에 따라 Recall은 증가하고 Precision은 감소하며, 추천집합의 크기가 감소함에 따라 Recall은 감소하고 Precision은

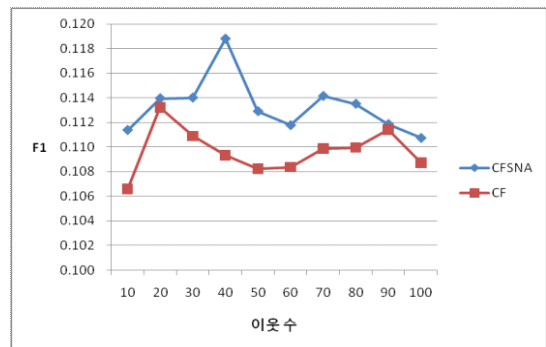
증가하게 된다. F1은 이러한 상충관계를 반영하여 측정하는 성능평가지표이다[1,6,7,10]. 검증데이터내의 각 고객별로 F1을 측정한 후 전체 134명 고객들의 평균 F1을 계산하여 추천 성능을 평가하였다.

V. 실험결과 및 분석

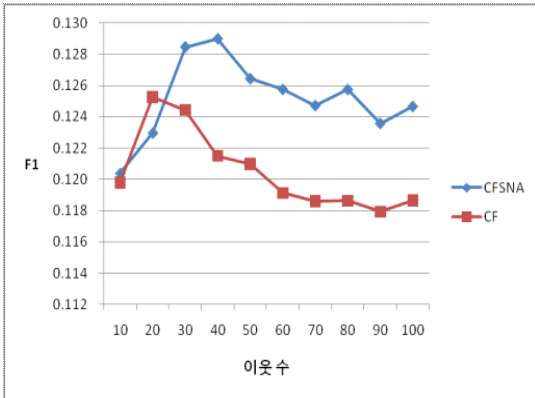
사회연결망을 활용한 상품추천 방안에 대한 세 가지 실험에서 연결정도 중심성을 이용하는 CFD의 경우 모든 상황에서 CF와 CFH에 비해 낮은 성능을 나타내었다. 기존 신규상품이나 신규고객을 위한 추천에서는 연결정도 중심성이 비교적 좋은 추천 정확성을 보였으나 이번 연구 실험에서는 낮은 정확성을 나타내었다.



(a) N=10



(b) N=15



(c) N=20

그림 3. CF와 CFH 실험결과

Fig. 3. Experimental Results of CF and CFH

연결정도 중심성을 활용하는 기존 연구[25]에서는 연결망을 구성 시에 식(3)을 이용하였다. 즉 반복 실험을 통해 높은 정확성을 보이는 유사도 임계값 ρ 를 찾고 이를 이용하여 연결망을 구성한 다음 연결정도 중심성을 사용하였다. 이 연구에서는 번거로운 반복 실험을 줄이고 유사도의 정보 손실을 막기 위해 식(4)와 같이 유사도를 직접 이용하여 연결망을 구성하였다. 이 실험결과를 통해 살펴보면 유사도를 직접 사용하는 경우에는 연결정도 중심성을 활용하는 것이 어려움을 알 수 있었다.

구조적 공백을 활용하는 경우(CFH)와 협업 필터링만을 사용하는 경우(CF)의 추천된 상품 수 N 에 따른 추천 성능 평가 결과를 그림 3에 제시하였다. 두 경우 ($N=10$ 이고 이웃고객 수=30인 경우와 $N=20$ 이고 이웃고객 수=20인 경우)만 제외하고 모든 경우에 CFH의 추천 성능이 CF에 비해 높음을 보여 주고 있다. 실험결과를 통해 분석해 보면, 협업 필터링 기반 추천시스템에서 이웃 선정과 상품 구매예측 점수 계산에 사회연결망의 구조적 공백 기법을 결합하여 활용함으로써 상품추천 성능을 높일 수 있는 것으로 판단된다.

VI. 결론

협업 필터링 기반 추천시스템에서 추천 성능은 전자상거래 환경이 확대되어 가고 있는 상황에서 고객 만족 및 매출증대에 기여할 수 있는 주요 이슈 중 하나이다. 최근 사회연결망 분석을 이용한 신규고객 및 신규제품 추천 시에 추천 성능이 높일 수 있음을 보이는 연구가 제시되고 있다. 본 연구에서는 사회연결망의 주요 분석 방법인 연결정도 중심성과 구조적 공백을 협업 필터링의 추천 프로세스에 반영하는 방안에 대해 제시하였다. MovieLens 영화추천 데이터에 실험 분석한 결과 구조적 공백 기법을 적절하게 활용할 경우 추천 성능을 높일 수 있음을 검증하였다. 연결정도 중심성의 경우 이번 실험에서는 추천 성능이 좋지 않게 나타났다. 이는 고객연결망이나 상품연결망을 구성하는 데 적절한 관계의 임계값을 반복실험을 통해 찾는 다른 연구에서 사용한 방법을 배제하고 유사도를 직접 사용하여 연결망을 구성한 때문으로 분석된다. 구조적 공백 위치를 차지한 개체는 정보획득 효과를 통해 연결망 내에서 영향력을 높일 수 있는 데, 이번 실험을 통해 협업 필터링의 이웃 선정 시 구조적 공백 값이 큰 고객을 포함시키고, 또한 상품 구매예측에서도 상품의 구조적 공백 값을 활용하는 것이 효과적임을 보여주었다.

이 연구에서는 영화추천 데이터를 이용하여 실험 검증하였는데, 다양한 전자상거래 데이터는 다양한 특성을 가지고 있으며, 따라서 사회연결망 분석 기법을 적용 시에 각 데이터 특성에 맞게 추가적인 실험분석을 하고 보정하는 과정이 필요할 것으로 판단한다. 또한 사회연결망 분석 기법은 매우 다양하며 지속적으로 발전적으로 확대되고 있다. 향후 이를 반영한 추가적인 연구는 추천시스템의 성능 향상에 크게 기여할 것으로 보인다.

참고문헌

- [1] 강부식, 유성진, 홍한국, "협력 여과기법의 추천 성능 개선을 위한 유사도 척도 비교", *Journal of the Korean Data Analysis Society*, Vol. 10, No. 2, pp. 2893-2906, 2008.
- [2] 조윤호, 방정해, "신상품추천을 위한 사회연결망 분석의 활용", *지능정보연구*, 제 15 권, 제 4 호, pp. 177-193, 2009.
- [3] 강현철, 한상태, 정병철, 신연주, "개인화를 위한 추천시스템 알고리즘에 관한 연구", *Journal of the Korean Data Analysis Society*, Vol. 6, No. 4, pp. 1043-1049, 2004.
- [4] 이희정, 홍태호, "클러스터링 기반 사례기반추론을 이용한 추천시스템 개발", *대한산업공학회/한국경영과학회 2004 춘계학술대회*, pp. 14-17, 2004.
- [5] 박종학, 조윤호, 김재경, "사회연결망: 신규고객 추천문제의 새로운 접근법", *지능정보연구*, 제 15 권, 제 1 호, pp. 123-140, 2009.
- [6] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Analysis of Recommendation Algorithms for E-Commerce", *Proceedings of the 2nd ACM Conference on Electronic Commerce*, pp. 158-167, October 2000.
- [7] 강부식, "사회연결망의 구조적 공백을 활용한 신규고객 웹 상품추천방법", *산업경제연구*, 제 23 권, 제 5 호, pp. 2371~2385, 2010.
- [8] 손동원, *사회네트워크 분석*, 경문사, 서울, 2002.
- [9] 김용학, *사회연결망분석*, 박영사, 서울, 2007.
- [10] 이재식, 박석두, "장르별 협력 필터링을 이용한 영화추천시스템의 성능 향상", *한국지능정보시스템학회논문지*, 제 13 권, 제 4 호, pp. 65-78, 2007.
- [11] R. S. Burt, *Structural Holes: The Social Structure of Competition*, Harvard University Press, 1992.
- [12] J. Herlocker, 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*, pp. 230-237, August 1999.
- [13] 조윤호, 김인환, "사회연결망 분석과 인공신경망을 이용한 추천시스템 성능 예측", *지능정보연구*, 제 16 권, 제 4 호, pp. 159~172, 2010.
- [14] W. D. Nooy, A. Mrvar, and V. Batagelj, *Exploratory Social Network Analysis with Pajek*, Cambridge University Press, 2005.

저자소개

강부식(Boo-Sik Kang)



1985년 경희대학교 산업공학과

(공학사)

1989년 한국과학기술원 산업공학과

(공학석사)

2000년 한국과학기술원 산업공학과

(공학박사)

1989년~2001년: 한국통신 연구원

2001년~현 재: 목원대학교 서비스경영학부 교수

※ 관심분야: 지능정보시스템, 데이터마이닝, 고객관계 관리, 서비스품질경영