지능정보연구 제17권 제4호 2011년 12월(pp.305~316)

추천시스템의 효과적 도입을 위한 소셜네트워크 분석

박종학
동양미래대학 e-비즈니스과
(jhpark@dongyang.ac.kr)

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

협업필터링은 다양한 분야에서 널리 활용되고 있지만 협업필터링의 추천 성능은 적용하는 기업의 비즈니스 형태나 발생하는 거래 데이터의 특성에 따라 다르게 나타나고 있다. 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 살 수 있다. 따라서 추천시스템 도입을 검토할 때 기업이 갖고 있는 데이터의 특성을 파악하고 이를 통해 추천시스템을 도입하는 것이 타당한지 사전에 예측할 수 있다면 불필요한 도입으로 인한 경제적 손실과 고객 만족도 저하를 막을 수 있을 것이다. 기존 연구에서는 협업필터링 추천 성과에 희박성, 우연성, 커버리지 등이 영향을 미칠 수 있다고 설명하고 있지만 이러한 요인들이 어떻게 얼마나 추천 성과에 영향을 미치는지, 요인들 간에 어떠한 상관관계가 있는지는 현재까지 구체적으로 밝혀진 바가 없다. 본 연구에서는 구매 트랜잭션으로부터 생성된 소셜네트워크로부터 밀도, 군집화계수, 집중도 등의 구조적 지표를 측정한 후 이들이 추천성과에 어떻게 영향을 미치는지 통계적 분석을 통해 실증적으로 규명한다. 이를 통해 협업필터링 추천시스템에 대한 도입 여부를 결정하고자 할 때 유용하게 사용될 수 있는 지침을 제공하고자 한다.

논문접수일 : 2011년 11월 18일 게재확정일 : 2011년 12월 20일
투고유형 : 학술대회우수논문 교신저자 : 조윤호

1. 서론

현재까지 추천시스템을 구현하기 위한 다양한 기법들이 개발되어 왔는데, 이중에서 협업필터링이 가장 성공적인 추천기법으로 알려져 있으며 Amazone.com, Netfix.com, CDNnow.com 등 수많은 기업들이 협업필터링을 통해 고객에게 추천서비스를 제공하고 있다(박종학 외, 2009; Su and Khoshgoftaar, 2009).

그럼에도 불구하고, 협업필터링의 추천 성능은 적용하는 기업의 비즈니스 형태나 발생하는 거래 데이터의 특성에 따라 다르게 나타나고 있다. 예를 들어 영화나 도서 판매 사이트에서는 추천의 정확도가 높지만, 상대적으로 의류 판매 사이트에서는 추천의 정확도가 낮은 것으로 알려져 있다(Huang et al., 2007). 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 살 수 있다. 따라

* 이 논문은 동양미래대학의 학술연구 지원에 의하여 연구되었음.

서 추천시스템 도입을 검토할 때 기업이 갖고 있는 데이터의 특성을 파악하고 이를 통해 추천시스템을 도입하는 것이 타당한지 사전에 예측할 수 있다면 불필요한 도입으로 인한 경제적 손실과 고객 만족도 저하를 막을 수 있을 것이다.

기존 연구에서는 협업필터링 추천 성과에 Sparsity, Gray sheep, Cold-start, Coverage, Serendipity 등이 영향을 미칠 수 있다고 설명하고 있다(Adomavicious and Tuzhilin, 2005; Sarwar et al., 2000; Su and Khoshgoftaar, 2009). 하지만 이러한 요인들이 어떻게 측정될 수 있는지, 어떻게 얼마나 성과에 영향을 미치는지, 요인 간에 어떠한 상관관계가 있는지는 현재까지 실증적으로 밝혀진 바가 없다.

협업필터링에서는 고객 간의 상품 선호도 또는 구매 연관성을 분석한 후 유사한 고객들을 묶어 상품을 추천하는 관계를 형성하는데, 이 관계를 그래프로 나타내어 네트워크를 형성하면 이 네트워크는 소셜네트워크가 된다(Ryu, et al., 2006). 즉, 협업필터링 추천시스템은 고객의 구매데이터를 분석하여 소셜네트워크를 인위적으로 생성한 후에 링크로 연결된 고객들(이웃고객)의 구매정보를 이용하여 특정 고객에게 상품을 추천할 수 있게 만든 일종의 소셜네트워크 시스템이라고 할 수 있다(조윤호, 김인환, 2010).

본 연구에서는 거래 데이터의 소셜네트워크 분석(Social Network Analysis : SNA)을 통해 추천 성과의 차이를 발생시키는 영향요인을 실증적으로 규명하고, 기업에서 협업필터링 도입 여부를 결정하고자 할 때 유용하게 사용될 수 있는 지침을 제공하고자 한다.

본 연구의 목적을 달성하기 위해 먼저 국내 유명 H 백화점의 거래 데이터를 기반으로 다양한 형태의 소셜네트워크를 구성하고 각 소셜네트워크로부터 여러 구조적 지표(독립변수)들을 측정하였다. 협업필터링 시스템을 구축하여 각 소셜네트워크의 추천 정확도(종속변수)를 계산한 후, 구조적 지표들이 추천 정확도에 미치는 영향을 실증적으로 분석하기 위해 3-way ANOVA 통계분석을 실시하였다.

2. 관련 연구

2.1 협업필터링에서의 추천성과 영향 요인

협업필터링 기반 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 고객-상품 행렬 구성, 유사 이웃 집단 탐색, 추천 상품 결정 단계로 구성된다(Sarwar et al., 2000). 협업필터링의 추천 성능은 적용하는 기업의 비즈니스 형태나 소유하고 있는 정보의 특성에 따라 다르게 나타난다(Huang and Zeng, 2005; Huang et al., 2007). 기존의 협업필터링 연구에서는 이러한 추천성과의 차이에 영향을 줄 수 있는 요인으로 다음과 같은 5가지의 요인을 언급하고 있다.

1) Sparsity : 기존의 많은 연구에서 상품에 대한 고객의 선호도 정보가 적으면 적을수록 추천시스템의 정확도가 떨어진다고 보고하고 있다(박종학 외, 2009; Herlocker et al., 2004; Su and Khoshgoftaar, 2009; Adomavicious and Tuzhilin, 2005). 이는 고객-상품 행렬이 희박 행렬(sparse matrix)이 되면 유사 이웃 집단을 탐색하는 과정에서 적은 수의 선호도 데이터를 사용하므로 고객간 유사도 측정 과정에서 신뢰성

이 떨어지기 때문이다.

2) Gray Sheep : 타 고객들과 다른 독특한 구매행태를 보이는 고객들을 Gray Sheep이라고 한다. 자신만의 독특한 구매를 하는 고객들은 다른 고객들과 유사한 선호도 정보가 없어 고객간 유사성을 찾기가 힘들다. 따라서 이러한 고객들이 차지하는 비율에 따라 추천 성능이 달라질 수 있다(Su and Khoshgoftaar, 2009).

3) Cold-start : 신규고객이나 구매 이력이 적은 고객들이 많을 수록 고객-상품 행렬은 희박하게 된다. 따라서 이러한 고객들이 많을수록 추천성능은 낮아지게 된다(Herlocker et al., 2004; Su and Khoshgoftaar, 2009; Adomavicious and Tuzhilin, 2005).

4) Coverage : 추천시스템의 커버리지는 추천을 하거나 예측할 수 있는 아이템의 범위를 말한다. 즉, 추천시스템이 전체 상품 중 사용자에 추천할 수 있는 상품의 비율을 의미한다(Herlocker et al., 2004). 커버리지가 넓어질수록 추천 대상 고객에게 더 많은 상품 선택의 기회를 제공하기 때문에 추천 성과가 높아질 수 있다(Herlocker et al., 2004).

5) Serendipity : 우연성은 고객이 과거에 구매하지도 전혀 고려하지도 않았었지만 의외로 높은 흥미를 가질 수 있는 전혀 다른 상품을 추천해 주는 것을 말한다(Herlocker et al., 2004; Murakami et al., 2008). 우연성이 높을수록 추천시스템의 커버리지 또한 넓어지게 되고, 이는 추천 성능의 향상으로 이어진다.

2.2 소셜네트워크 분석과 구조적 지표

지난 수 십 년 동안 소셜네트워크의 구조를 측정하기 위한 개념 또는 지표들이 다양하게 개발되어 왔다. 먼저, 네트워크의 결속(cohesion)을 측정하기 위한 측정지표로 포괄성, 연결정도, 밀도, 군집화계수, 이행성 등이 있다(Frank and Harary, 1982; Wasserman and Faust, 1994; Watts, 1999; 손동원, 2002; 김용학, 2003). 네트워크에서 중심에 위치하는 정도와 중심에 집중된 정도를 나타내는 측정지표로는 연결정도(degree) 중심성, 근접(closeness) 중심성, 매개(betweeness) 중심성, 집중도 등이 있다(Freeman, 1979; Bonachich, 1987; 손동원, 2002). 소셜네트워크를 구성하는 하부 집단을 규명하기 위한 측정지표들로는 파당(faction), 결속집단(clique) 등이 있다(Seidman and Foster, 1978; Amorim, et al., 1992). 또한 소셜네트워크에서 역할과 위치에 대한 구조적 등위성을 위한 측정지표로는 CONCOR, STRUCTURE 등이 있다(Breiger et al., 1975; Burt, 1991; 손동원, 2002).

본 연구에서는 이러한 지표들 중에 제 2.1절에서 다룬 추천성과 영향요인들과 관련 있는 밀도, 군집화계수, 집중도를 이용하여 추천 성과의 차이를 규명하고자 한다.

밀도는 한 네트워크의 참여자 간의 가능한 모든 관계 중에서 실제로 맺어진 관계 수의 비율로 정의된다(김용학, 2003; 손동원, 2002). 식 (1)에서 n은 네트워크 전체 노드의 수이고 k는 실제 연결된 관계의 수를 나타낸다.

$$밀도 = \frac{k}{n(n-1)/2} \qquad (1)$$

협업필터링 관점에서 밀도는 고객 사이에 얼마나 많은 유사한 구매패턴 관계를 가지고 있는가로 해석될 수 있다. 따라서 밀도의 증가는 고객-상품 행렬에서 비어있는 셀의 감소 즉, 희박성의 개선을 의미한다고 볼 수 있다. 결론적으로 밀도가 증가할

수록 추천 정확도는 높아질 것이다.

군집화계수는 네트워크 내의 3명의 참여자 a, b, c가 있고 a와 b, a와 c사이에 관계가 있을 때 b와 c가 관계를 가질 가망성을 말한다. 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 (2)과 같이 네트워크의 각 노드의 입장에서 존재하는 삼각관계의 수를 삼자관계의 수로 나누어 계산한다. 각 노드의 군집화계수를 평균한 것이 네트워크 수준의 군집화계수이다. 전체 네트워크의 군집화계수는 식 (3)을 통하여 측정된다(Schank and Wagner, 2005).

$$c(a) = \frac{a\text{의 삼각관계의 수}}{a\text{의 삼자관계의 수}} \quad (2)$$

$$CC = \frac{1}{|V|}\sum_{a \in V} c(a), \quad (3)$$

단, V는 연결정도가 2이상인 노드의 집합

군집화계수는 추천대상 고객과 이웃 고객들과의 관계 강도를 나타낸다고 볼 수 있다. 이웃간의 관계가 많이 형성되어 있다는 것은 그만큼 이웃들 간의 구매 패턴이 유사하다는 것을 의미한다. 따라서 추천의 커버리지는 감소하고 추천의 우연성 또한 감소할 수 있다. 반대로 군집화 계수가 낮을수록 이웃들의 구매 패턴은 다양하다고 볼 수 있으며 이에 따라 커버리지와 우연성이 증대될 것이다. 결론적으로 군집화계수가 낮아질수록 우연성 추천을 포함한 추천 커버리지가 증대되고 추천의 성과는 향상된다고 볼 수 있다.

집중도는 중심성이 높은 참여자에 얼마나 관계가 집중되어 있는지를 나타낸다(손동원, 2002; Bonacich, 1987; Freeman, 1979). 집중도는 0에서 1사이의 값을 가지는데 네트워크 구조가 관계를 많이 가지는 참여자에게 집중될수록 값은 1에 가까워진다. 극단적으로 스타형 구조의 네트워크는 하나의 참여자만이 다른 모든 참여자와 관계를 형성하는데 이러한 구조를 중앙집중형 네트워크 구조라고 할 수 있고 집중도 값은 1을 나타내게 된다. 반대로 원형의 네트워크 구조가 되면 모든 참여자들이 동일한 관계를 가지게 되며 집중도 값은 0을 갖게 된다(손동원, 2002; Freeman, 1979; Scott, 2000; Wasserman and Faust, 1994). 네트워크 전체 참여자의 수를 n이라고 할 때, 연결정도 집중도는 식 (4)와 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).

$$NC_D = \frac{\sum_{i=1}^{n}[C_D(P^*) - C_D(P_i)]}{max\ \sum_{i=1}^{n}[C_D(P^*) - C_D(P_i)]} = \frac{\sum_{i=1}^{n}[C_D(P^*) - C_D(P_i)]}{[(n-1)(n-2)]} \quad (4)$$

집중도가 높아 핵심 고객에 의존적이면 소수의 핵심 고객에 대한 추천 정확도는 많은 변두리 고객들의 의견을 반영할 수 있어 높아지겠지만 다수의 변두리 고객들은 소수의 핵심고객 정보만을 가지고 추천을 받기 때문에 추천 커버리지가 감소되어 결국 추천의 정확도는 떨어질 것이다.

3. 연구 방법

3.1 데이터 수집

본 실험에서는 조윤호, 김인환(2010)의 연구에서 사용한 H 백화점 구매 데이터를 이용하였다. 실험 데이터는 2000년 5월 1일부터 2001년 4월 30일까지 1년 동안 50,000명의 고객들이 4,038개의 상품에 대해 일으킨 1,660,814건의 구매 트랜잭션으로부터 표본 추출된 다양한 형태의 소셜네트워크를 구축할 수 있는 396개의 데이터 셋으로 구성되어 있다.

3.2 독립변수 측정

조윤호, 김인환(2010)이 제안한 구매 트랜잭션으로부터 소셜네트워크를 구축하는 방법(<그림 1> 참조)에 따라 396개의 데이터 셋으로부터 각각 소셜네트워크를 구축하고, UCINET6.0을 통하여 각 네트워크에 대한 밀도, 집중도 그리고 군집화계수를 측정하였다.

3.3 종속변수 측정

협업필터링 추천성과를 측정하기 위하여 널리 활용되고 있는 User-based 알고리즘(Sarwar et al., 2000; Su and Khoshgoftaar, 2009)을 적용하여 추천시스템을 구축한 후, 구축된 추천시스템을 396개의 데이터 셋 각각에 적용하여 추천 정확도를 측정하였다. 추천 정확도를 측정하기 위한 지표로 아래와 같이 계산되는 *F1-measure*를 사용하였다(Sarwar et al., 2000; Herlocker et al., 2004; 박종학 외, 2009).

$$F1 = \frac{recall \times precisiion}{(recall + precisiion)/2} \qquad (5)$$

3.4 실험 설계

밀도, 군집화계수, 집중도의 고저에 따라 추천 정확도(F1)가 달라지는지 그리고 각 지표 간에 상호작용효과가 있는지 검증하기 위하여, 본 연구의 실험은 2개의 밀도 수준(평균보다 높은 집단 vs. 낮은 집단)×2개의 군집화계수 수준(평균보다 높은

<그림 1> 소셜네트워크 구축 프로세스

집단 vs. 낮은 집단)×2개의 집중도 수준(평균보다 높은 집단 vs. 낮은 집단)의 3요인 팩토리얼 실험(three-factor factorial experiment)으로 설계되었다. 3-way ANOVA 분석을 수행하여 3가지 지표의 주효과(main effect), 2-way 상호작용효과(interaction effect), 그리고 3-way 상호작용효과가 각각 존재하는지 살펴보았다.

4. 연구 결과

각 집단의 평균 추천 정확도는 <표 1>과 같다. 밀도, 군집화계수, 집중도를 독립변수로 하고 F1을 종속변수로 두어 3-way ANOVA 분석을 실시한 결과, <표 2>에 보는 바와 같이 각 구조적 지표의 주효과와 두 지표 간의 2-way 상호작용효과가 유의하게 나타났다($p < 0.01$). 반면 세 지표간의 3-way 상호작용효과는 없는 것으로 분석되었다.

밀도와 집중도가 높을수록 추천 정확도가 높아지는 반면, 군집화계수가 높아지면 추천 정확도는 떨어진다(<그림 2>). 집중도의 경우 값이 높으면

<표 1> 구조적 지표에 따른 추천 정확도

SN Patterns	F1	Group No.
dDensity ↑, dCluster ↑, dCentral ↑	.026579	1
dDensity ↑, dCluster ↑, dCentral ↓	.027696	2
dDensity ↑, dCluster ↓, dCentral ↑	.028083	3
dDensity ↑, dCluster ↓, dCentral ↓	.026990	4
dDensity ↓, dCluster ↑, dCentral ↑	.007823	5
dDensity ↓, dCluster ↑, dCentral ↓	.006457	6
dDensity ↓, dCluster ↓, dCentral ↑	.021111	7
dDensity ↓, dCluster ↓, dCentral ↓	.010843	8

주) dDensity : 밀도, dCluster : 군집화계수, dCentral : 집중도.

추천 커버리지가 감소되어 추천 정확도가 떨어질 것으로 예상했으나 이와 상반된 실험 결과가 나왔다. 어느 정도의 집중화는 성능 향상에 긍정적인 영향을 미친다고 판단된다.

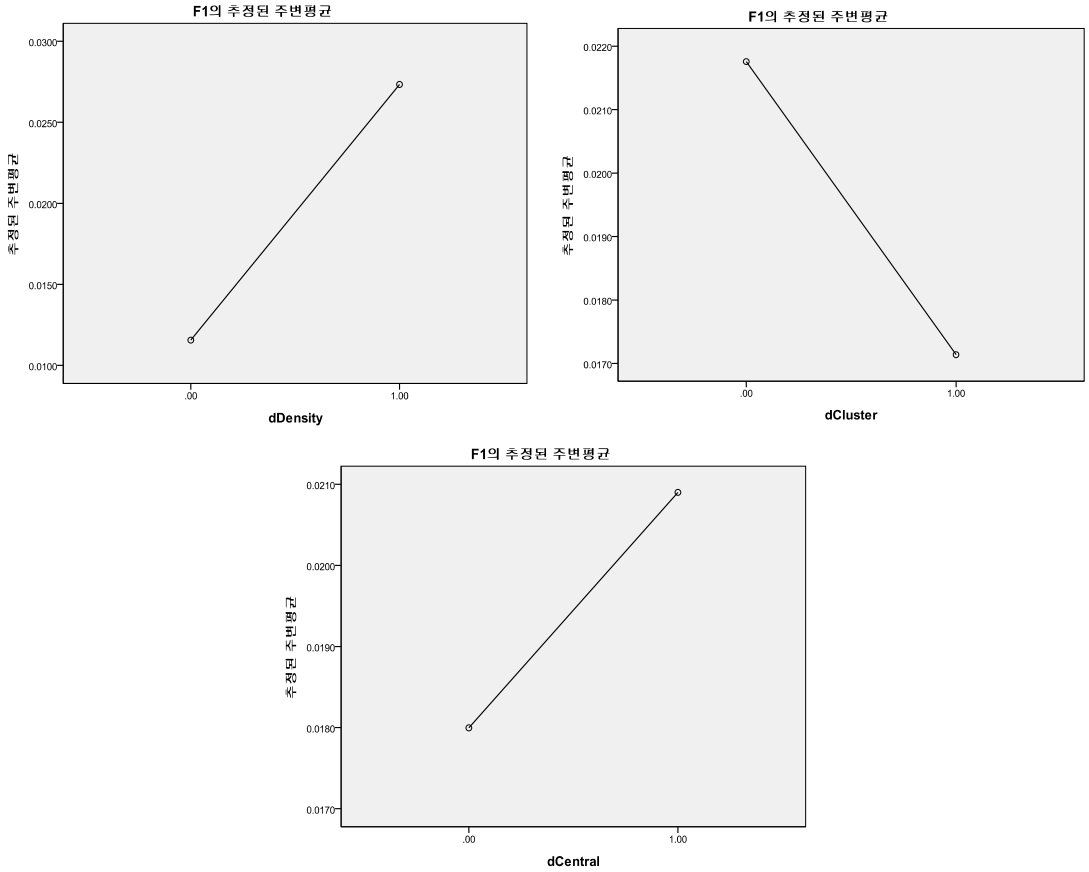
특히 밀도가 추천 정확도에 미치는 영향은 다른 두 지표의 영향 보다 훨씬 큰 것으로 나타났다. <표 1>에서 보는 바와 같이 밀도가 낮은 집단(No. 5-8)보다 밀도가 높은 집단(No.1-4)은 추천 성능

<표 2> 3-way ANOVA 분석결과

종속변수 : F1

소스	제 III유형 제곱합	자유도	평균제곱	F	유의확률
수정 모형	.028[a]	7	.004	65.604	.000
절편	.093	1	.093	1503.121	.000
dDensity	.015	1	.015	247.364	.000
dCentral	.001	1	.001	8.369	.004
dCluster	.001	1	.001	21.190	.000
dDensity×dCentral	.001	1	.001	8.439	.004
dDensity×dCluster	.001	1	.001	17.686	.000
dCentral×dCluster	.000	1	.000	7.669	.006
dDensity×dCentral×dCluster	.000	1	.000	2.781	.096
오차	.024	388	6.162E-5		
합계	.154	396			
수정 합계	.052	395			

주) [a] R 제곱 = .542(수정된 R제곱 = .534).



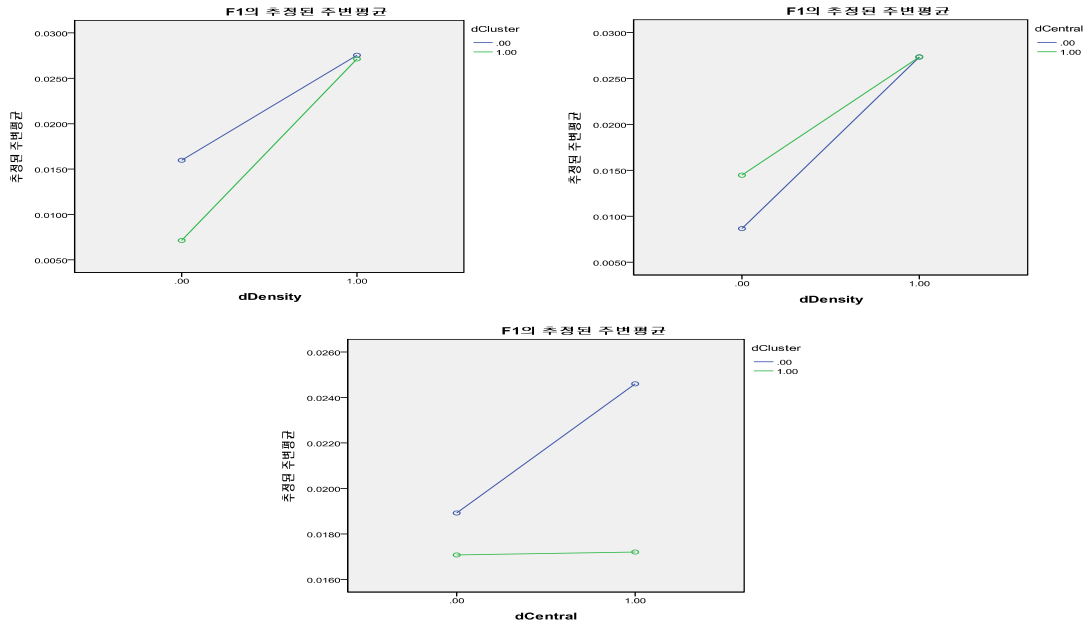
<그림 2> 각 지표의 주효과

이 훨씬 높으면서 서로 거의 비슷한 결과를 보이고 있다. 이는 고객별 거래 데이터가 많지 않거나 취급 상품의 종류가 너무 다양한 기업의 경우(즉, 고객-상품 행렬이 희박한 경우) 추천시스템 도입을 도입하더라도 소기의 성과를 얻기 어려울 수 있다는 것을 의미한다.

하지만, <그림 3>에서 보는 것과 같이 밀도가 낮은 집단에서도 군집화계수가 낮을 경우(No.7-8)에는 그렇지 않은 경우(No.5-6)보다 성능이 높게 나타남을 알 수 있다. 특히 군집화계수가 낮고 집중도가 높을때(No.7)는 밀도가 높은 집단들(No.1-4)과 비교해도 성능의 차이가 크게 나지 않는다. 이는 희박한 트랜잭션 데이터를 갖고 있는 기업의 경우에도 구매 트랜잭션 데이터의 다른 특성에 따라(즉, 군집화계수가 낮으면서 집중도가 높으면) 추천시스템 도입이 타당할 수 있다는 것을 암시하는 결과이다.

5. 결론

본 연구에서는 구매 트랜잭션 데이터에 대한 소셜네트워크 분석을 통해 협업필터링 추천 성과의



<그림 3> 각 지표간 2-way 상호작용효과

차이를 발생시키는 영향 요인으로 밀도, 군집화계수, 집중도가 있음을 실증적으로 규명하였다. 또한 이들 영향 요인들 간에 상호작용효과가 존재함도 알 수 있었다. 이를 통해 기업에서 추천시스템 도입을 검토할 때 추천시스템을 도입하는 것이 타당한지 사전에 검토할 수 있는 지침을 제공하였다. 본 연구의 결과를 통해 추천시스템의 불필요한 도입으로 인한 경제적 손실과 고객 만족도 저하를 막을 수 있을 것으로 기대한다.

참고문헌

김용학, 사회연결망 분석, 박영사, 2003.

손동원, 사회 네트워크 분석, 경문사, 2002.

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

조윤호, 김인환, “사회연결망분석과 인공신경망을 이용한 추천시스템 성능 예측”, *지능정보연구*, 16권 4호(2010), 159~172.

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.

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

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.

Herlocker, J. L., J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender systems", *ACM Transactions on Information Systems*, Vol.22, No.1(2004), 5~53.

Huang, Z. and D. Zeng, "Why Does Collaborative Filtering Work? Recommendation Model Validation and Selection by Analyzing Random Bipartite Graphs", Proceedings of 15th 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.

Murakami, T., K. Mori, and R. Orihara, "Metrics for evaluating the serendipity of recommendation lists", *Lecture Notes in Computer Science*, Vol.4914(2008), 40~46.

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, 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, No.4(2009).

Wasserman, S. and K. Faust, "Social network analysis : Methods and application", New York : Cambridge University Press, 1994.

Watts, D. J., "Small worlds, Princeton", NJ : Princeton University Press, 1999.

Abstract

Social Network Analysis for the Effective Adoption of Recommender Systems

Jong Hak Park* · Yoonho Cho**

Recommender system is the system which, by using automated information filtering technology, recommends products or services to the customers who are likely to be interested in. Those systems are widely used in many different Web retailers such as Amazon.com, Netfix.com, and CDNow.com. Various recommender systems have been developed. Among them, Collaborative Filtering (CF) has been known as the most successful and commonly used approach. 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. However, 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 in advance whether the performance of CF recommender system is acceptable or not is practically important and needed. In this study, we propose a decision making guideline which helps decide whether CF is adoptable for a given application with certain transaction data characteristics.

Several previous studies reported that sparsity, gray sheep, cold-start, coverage, and serendipity could affect the performance of CF, but the theoretical and empirical justification of such factors is lacking. Recently there are many studies paying attention to Social Network Analysis (SNA) as a method to analyze social relationships among people. SNA is a method to measure and visualize the linkage structure and status focusing on interaction among objects within communication group. CF analyzes the similarity among previous ratings or purchases of each customer, finds the relationships among the customers who have similarities, and then uses the relationships for recommendations. Thus CF can be modeled as a social network in which customers are nodes and purchase relationships between customers are links. Under the assumption that SNA could facilitate an exploration of the topological properties of the network structure that are implicit in transaction data for CF recommen-

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

dations, we focus on density, clustering coefficient, and centralization which are ones of the most commonly used measures to capture topological properties of the social network structure. While network density, expressed as 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. Centralization reflects the extent to which connections are concentrated in a small number of nodes rather than distributed equally among all nodes. We explore how these SNA measures affect the performance of CF performance and how they interact to each other.

Our experiments used sales transaction data from H department store, one of the well-known department stores in Korea. Total 396 data set were sampled to construct various types of social networks. The dependant variable measuring process consists of three steps; analysis of customer similarities, construction of a social network, and analysis of social network patterns. We used UCINET 6.0 for SNA. The experiments conducted the 3-way ANOVA which employs three SNA measures as dependant variables, and the recommendation accuracy measured by F1-measure as an independent variable. The experiments report that 1) each of three SNA measures affects the recommendation accuracy, 2) the density's effect to the performance overrides those of clustering coefficient and centralization (i.e., CF adoption is not a good decision if the density is low), and 3) however though the density is low, the performance of CF is comparatively good when the clustering coefficient is low. We expect that these experiment results help firms decide whether CF recommender system is adoptable for their business domain with certain transaction data characteristics.

Key Words : Social Network, Collaborative Filtering, Density, Clustering Coefficient, Centralization

저 자 소 개

박종학
현재 동양미래대학 e-비즈니스과 부교수로 재직 중이다. 서울대학교 계산통계학과(전산학전공)에서 학사, KAIST 경영정보공학과에서 석사를 취득하였으며, KAIST 테크노경영대학원에서 박사과정을 수료하였다. LG전자(주)에서 5년간 주임연구원으로 재직한 바 있으며 주 연구 관심분야는 e-비즈니스, 데이터마이닝 등이다.

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