ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2014 June: 20(2): 137~148 http://dx.doi.org/10.13088/jiis.2014.20.2.137

소셜 네트워크 분석 기법을 활용한 협업필터링의 특이취향 사용자(Gray Sheep) 문제 해결

김민성
연세대학교 경영대학
(minsung721@gmail.com)

임 일
연세대학교 경영대학
(il.im@yonsei.ac.kr)

상품 검색시간의 단축과 쇼핑에 투입되는 노력의 감소 등, 온라인 쇼핑이 주는 장점에 대한 긍정적인 인식이 확산되면서 전자상거래(e-commerce)의 중요성이 부각되는 추세이다. 전자상거래 기업들은 고객확보를 위해 다양한 인터넷 고객관계 관리(eCRM) 활동을 전개하고 있는데, 개인화된 추천 서비스의 제공은 그 중 하나이다. 정확한 추천 시스템의 구축은 전자상거래 기업의 성과를 좌우하는 중요한 요소이기 때문에, 추천 서비스의 정확도를 높이기 위한 다양한 알고리즘들이 연구되어 왔다. 특히 협업필터링(collaborative filtering: CF)은 가장 성공적인 추천기법으로 알려져 있다. 그러나 고객이 상품을 구매한 과거의 전자상거래 기록을 바탕으로 미래의 추천을 하기 때문에 많은 단점들이 존재한다. 신규 고객의 경우 유사한 구매 성향을 가진 고객들을 찾기 어렵고 (Cold-Start problem), 상품 수에 비해 구매기록이 부족할 경우 상관관계를 도출할 데이터가 희박하게 되어(Sparsity) 추천성능이 떨어지게 된다. 취향이 독특한 사용자를 뜻하는 'Gray Sheep'에 의한 추천성능의 저하도 그 중 하나이다.

이러한 문제인식을 토대로, 본 연구에서는 소셜 네트워크 분석기법 (Social Network Analysis: SNA)과 협업필터링을 결합하여 데이터셋의 특이 취향 사용자 (Gray Sheep) 문제를 해소하는 방법을 제시한다. 취향이 독특한 고객들의 구매 데이터를 소셜 네트워크 분석지표를 활용하여 전체 데이터에서 분리해낸다. 그리고 분리한 데이터와 나머지 데이터인 두 가지 데이터셋에 대하여 각기 다른 유사도 기법과 트레이닝 셋을 적용한다. 이러한 방법을 사용한 추천성능의 향상을 검증하기 위하여 미국 미네소타 대학 GroupLens 연구팀에 의해 수집된 무비렌즈 데이터(http://movielens.org)를 활용하였다. 검증결과, 일반적인 협업필터링 추천시스템에 비하여 이 기법을 활용한 협업필터링의 추천성능이 향상됨을 확인하였다.

주제어 : 협업필터링, 소셜 네트워크 분석, Gray Sheep Problem

논문접수일 : 2014년 6월 15일 논문수정일 : 2014년 6월 19일 게재확정일 : 2014년 6월 22일
투고유형 : 국문급행 교신저자 : 임일

1. 서론

IT발달로 인한 전자상거래 (e-commerce) 의 급격한 팽창과 데이터 마이닝 기법의 발달로 온라인상에서의 효율적인 고객관리를 위한 인터넷 고객관계 관리(eCRM)에 대한 연구가 많이 이루어지고 있다. 온라인쇼핑의 가장 큰 장점은 편리함인데 편리함에는 소비자가 원하는 것을 구매하기 위한 검색시간, 노력의 감소가 포함된다. 따라서 두 가지를 줄이는 것이 온라인시장에서 성공의 열쇠이다(Hung, 2005). 온라인쇼핑 고객에게 편리함을 제공하기 위하여 전자상거래

* 이 논문은 2013년 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2013S1A3A2055050)

(e-commerce) 기업들이 행하는 가장 대표적인 인터넷 고객관계 관리 (eCRM) 활동이 개인화된 추천시스템이다. 개인화된 추천시스템은 온라인상에서 고객들에게 일대일 마케팅(One-to-one marketing) 제공을 가능케 하여 기업성과에 큰 영향을 준다 (Hung, 2005; Sarwar at al., 2000).

협업필터링 추천기법은 과거 구매 기록을 통해 고객의 구매성향을 판단하고 구매성향이 가장 높은 상관관계를 보이는 다른 고객의 구매기록에서 제품을 추천하는 구현방식이다 (Sarwar et al., 2000; Im and Hars, 2007). 이러한 협업필터링 추천기법은 성공적인 추천기법으로 알려져 있기 때문에 Amazon과 Netflix 등 여러 기업에서 이 기법을 사용하여 고객에게 상품을 추천하고 있다(Herlocker et al., 2002)

그러나 협업필터링 기술은 과거의 구매기록을 바탕으로 추천을 하기 때문에 데이터 희박성(sparsity), 독특한 취향의 고객(gray sheep), 신규고객 추천의 문제(cold-start problem), 그리고 우연의 추천(serendipity) 등과 같은 단점이 존재한다 (Ghazanfar and Adam, 2014; Sarwar et al, 2000; Su and Khoshgoftaar, 2008). 이는 협업필터링 추천시스템의 성능 차이의 원인이 되기도 한다.

이러한 기존의 협업필터링 기술이 가지고 있던 대부분의 문제점들은 거래 데이터(transaction data)에서 다른 검증 없이 고객들의 구매행위를 뽑아낸 데이터셋의 특성에 기인한 것이다. Im and Hars (2007)는 이 같은 데이터셋 특성의 중요성을 주목하여 도메인별 추천시스템의 성능차이를 이질적(heterogeneous), 동질적(homogeneous) 집단간의 데이터셋 특성으로 설명하였다. 또한, 소셜 네트워크 관점에 기반하여 소셜 네트워크 분석지표를 활용하여 데이터셋 특성에 따른 추천성능을 예측하고자 하는 연구들이 있었다 (Ahn et al., 2012; Park and Cho, 2011). 이들 연구에 의하면 데이터셋 네트워크의 밀도, 군집화 계수, 집중도 정도를 보고 추천정확도의 예측이 가능하다.

협업필터링의 단점을 보완하고 추천성능을 향상시키고자 하는 또 다른 측면에서의 접근으로서 내용 기반 필터링 방식과 결합한 하이브리드 협업필터링에 대한 연구가 있었고 (Barragans et al., 2010; Konstan et al., 1998) 최근에는 소셜 네트워크분석 기법을 활용한 협업필터링 추천시스템에 대한 연구가 활발히 진행되고 있다. 구체적으로, 중심성(centrality)이 높은 고객이 구매한 상품을 신규고객에게 추천하는 방법으로 신규고객 추천 문제를 해소한 연구가 있고(Shin et al., 2012) 데이터 희박성(data sparsity) 문제를 해소하고자 유사한 사용자 그룹을 찾아 클러스터링하는 과정에서 소셜 네트워크 분석을 활용한 연구가 있다 (Pham et al., 2011).

그러나 협업필터링의 단점을 보완하고자 하는 연구가 활발히 진행되어온 반면, 데이터의 특성에 주목하여 데이터 구조 개선과 자체튜닝을 통한 성능향상에 관심을 기울인 연구는 거의 이루어지지 않았다. 데이터 튜닝을 실시하지 않은 데이터셋을 협업필터링 분석에 그대로 사용할 경우, 추천성능을 저하시키는 '독특한 취향의 고객들(gray sheep)'의 정보에 의한 왜곡된 결과가 발생할 수 있다. 따라서 독특한 취향의 고객들로 인한 '데이터셋 오염'과 이에 따른 분석의 왜곡 문제를 방지하고 추천성능 향상을 위해서는 데이터셋 튜닝을 실시하여야 한다.

본 연구는 소셜 네트워크 분석(Social Network Analysis: SNA) 기법을 활용하여 데이터셋을 튜닝하는 협업필터링 (Collaborative Filtering: CF)

추천시스템을 개발하고 시뮬레이션을 통해 이를 검증하였다. 보다 상세하게는 고객들의 거래데이터에서 독특한 취향의 고객들(isolate nodes)을 소셜 네트워크 분석 지표를 활용하여 분리해 내고, 이를 통한 거래 데이터셋 자체의 튜닝으로 추천 시스템의 성능을 향상시키는 방법에 관한 것이다.

2. 이론적 배경

2.1 추천 알고리즘에 관한 연구

Goldberg(1992)에 의해 협업필터링 추천시스템이 제안된 이후, 이 시스템은 기업현장에서 가장 성공적인 추천기법으로 사용되고 있다. 그러나 과거 데이터를 바탕으로 미래 추천을 생성하는 알고리즘으로 인하여 여러 단점이 존재하였고, 2000년대 중반까지 이 문제를 극복하기 위한 하이브리드 알고리즘에 관한 연구들이 함께 이루어졌다. Claypool (1999)은 신규고객이나 새로운 상품에 대한 추천이 어렵다는 문제점(cold-start problem)을 해결하기 위하여 아이템의 내용을 추천에 활용하는 내용 기반 필터링(Content-based Filtering; CB)과 결합하는 방법을 제안하였다. Lee (2003)는 인구통계학적 정보를 협업필터링 추천에 활용하였다. 그러나 추가적인 정보를 활용하는 방식은 이의 확보를 위해 많은 비용과 시간의 투입이 필요하다는 단점이 있다. 추가적인 정보를 활용하는 알고리즘 외에 Konstan and Herlocker (1997)은 뉴스들을 클러스터링 방식을 통해 협업필터링의 데이터 희박성 문제를 해결하고자 하였다. 최근에는 협업 필터링에 클러스터링과 내용 기반 필터링 알고리즘을 결합하여 독특한 취향의 고객(gray sheep)들에 의한 문제점을 해결하고자 한 연구도 진행되었다(Ghazanfar and Adam, 2014). 클러스터링 알고리즘으로 독특한 고객을 구분해내고 이들에게 아이템의 내용을 활용하여 추천을 생성하는 방식이다. 그러나 이 방법에는 많은 컴퓨팅파워와 계산시간이 추가로 필요하다는 문제점이 존재한다. 클러스터링 알고리즘은 고객의 수가 늘어남에 계산시간이 기하급수적으로 늘어난다.

2000년대 중반 이후부터는 알고리즘 차원의 연구에서 벗어나 근본적으로 추천성과의 차이가 나타나는 원인을 규명하고자 한 연구가 이루어졌다. 데이터 셋 체계의 특성과 사용자 검색모드 관점에서 성능 차이를 분석을 한 연구가 있었고 (Im and Hars, 2007), 데이터 셋의 성격을 소셜 네트워크 분석지표를 통해 알아보고 성능을 예측하고자 한 연구가 있었다 (Park and Cho, 2011). 최근에는 데이터셋을 네트워크 체계의 관점에서 보고 소셜 네트워크 분석기법을 활용한 추천시스템 연구가 활발히 진행되고 있다.

2.2 소셜 네트워크 분석

소셜 네트워크 분석 기법은 사회과학뿐 아니라 물리학, 생물학 등 다양한 분야에서 응용되고 있다. 네트워크가 사람들 사이의 네트워크 뿐 아니라 신경망과 같은 생물학적 네트워크, 전력망에서 볼 수 있는 기술적 네트워크 등 다양한 형태로 존재하기 때문이다 (Dorogovtsev and Mendes, 2002; Newman, 2003).

인터넷과 컴퓨팅 기술이 발전하면서부터는 수백만 개의 노드를 가진 대용량 네트워크 데이터 분석이 가능해지게 되었고 (Kim et al., 2009; Newman, 2003), 이에 따라 네트워크 분석 기법

의 활용 범위는 더욱 광범위해졌다. 따라서 데이터의 범위가 넓은 인터넷 데이터의 네트워크 분석이 이전보다 용이해졌고, 네트워크 체계를 구성하는 노드들의 행위를 규명한다는 목적에서 전자상거래의 추천시스템 연구에도 활용되기 시작하였다.

중앙성(centrality)은 소셜 네트워크 분석에서 쓰이는 대표적인 지표 중 하나로 연결 정도 중앙성(degree centrality), 인접 중앙성(closeness centrality), 사이 중앙성(betweenness centrality)이 있으며 이 세 가지는 각기 다른 '중심'을 측정한다(Kim, 2013; Sohn, 2008). 본 연구에서는 사용자와 선호하는 영화간의 네트워크에서 특이한 취향의 사용자(gray sheep), 즉 다른 노드와의 연결이 되지 않았거나 링크(link) 수가 작은 노드를 분류해내기 위하여 다른 노드와의 연결된 정도를 의미하는 연결 정도 중앙성 지표를 사용한다(Sohn, 2008). 연결 정도 중앙성은 한 노드에 직접 연결되어 있는 노드들의 합으로 간단히 측정이 가능하며 구하는 식은 아래와 같다.

$$c_i = \sum_{j=1}^{n}(Z_{ij}+Z_{ji}) / \sum_{i=1}^{n}\sum_{j=1}^{n}(Z_{ij}) \text{ (단, } 0 \le c_i \le 1) \quad (1)$$

소셜 네트워크 지표를 이용하여 추천 데이터셋을 오염시키는 노드들을 분류해내고, 이를 통해 데이터를 튜닝하는 알고리즘에 관한 자세한 설명은 다음 절에 기술한다.

3. 연구방법

3.1 데이터셋

본 실험에서는 미네소타대학(University of Minnesota)의 GroupLens Research Project팀에 의해 수집된 MovieLens 데이터를 사용하였다. 데이터는 MovieLens 웹사이트(http://movielens.org)를 통해 1997년 9월 19일부터 1998년 4월 22일까지 7개월 간 수집되었다. 총 1,682개의 영화에 대해 943명의 고객들이 리커트 5점 척도로 매긴 100,000 건의 평가 데이터이다.

3.2 알고리즘

실험을 위하여 R의 tnet과 Recommanderlab 패키지를 사용하여 추천시스템을 구축하고 시뮬레이션을 진행하였다. 추천알고리즘의 적용 단계에서는 일반적으로 많이 사용하는 사용자기반 협업필터링 알고리즘과 유사도 기법으로 코사인 유사도(Cosine similarity)와 베스트셀러 추천(popular similarity)을 사용하였다.

본 연구에서는 고객의 거래데이터에 협업필터링 알고리즘을 적용하기 이전에, 독특한 취향의 고객들(gray sheep)을 분리하기 위하여 소셜 네트워크 분석의 연결 정도 중앙성(degree centrality) 지표를 이용하였다. 노드들을 연결 정도 중앙성의 값 순으로 정리하고, 시뮬레이션을 통하여 결정된 기준 값을 기준으로 연결 정도 중앙성이 낮은 데이터셋과 그렇지 않은 나머지 데이터셋으로 분리하였다. 이 각각의 데이터셋에 대하여 최고의 성과를 내는 다른 유사도 기법(Similarity Measure)기법과 다른 트레이닝셋(training set)을 적용한다.

자세한 순서는 다음과 같다.

Step 1: 사용자와 그 사용자가 선호하는 영화들로 구성되어 있는 2원 연결망 (2-mode network)을 사용자-사용자 네트워크로 구성되는 1원 연결망(1-mode network)으로

투영(projection) 한다.

Step 2: 소셜 네트워크 분석지표인 연결 정도 중앙성(degree centrality)을 활용하여 하위 값을 가지는 일정 노드들을 분리한다. 단, 분리 기준이 되는 연결 정도 중앙성의 값은 시뮬레이션 방식을 사용하여 추천성능이 최적일 때(하위 값을 가지는 노드가 제외된 데이터셋의 추천정확도가 최대일 때)의 값을 기준점으로 한다.

Step 3: 연결 정도 중앙성의 값이 하위인 일정 노드들을 분리해 내고 나머지 데이터셋에 대해 일반적인 시뮬레이션 방식을 통해 최상의 성과를 내는 협업필터링(CF)의 유사도 기법(코사인)을 찾아 추천을 시행한다.

Step 4: 분리된 하위 노드들의 데이터셋에서는 독특한 취향의 고객들이 몰려있기 때문에 위에서 구한 동일한 유사도 기법을 적용하면 추천아이템이 생성되지 않거나 생성되더라도 정확도가 낮을 가능성이 많다. 따라서 이 노드들에 대해서는 노드를 분리하기 전의 모든 고객들이 포함된 전체 데이터를 트레이닝셋(training set)으로 지정하고, 독특한 고객들을 의미하는 하위 노드들을 테스트셋(test set)으로 하여 인기항목 추천(popular items) 방법으로 추천을 실시한다.

3.3 추천성능 평가방법

본 논문에서 사용한 소셜 네트워크 분석(SNA) 기법을 활용한 협업필터링 추천시스템의 경우, 연결 정도 중앙성 값에 의해 분리된 2가지 데이터 셋의 추천성과가 각각 다르다. 따라서 공정한 비교를 위하여 노드들의 숫자비율에 따라 이 2가지 데이터 셋의 추천성능의 가중평균을 구하고, 이를 일반적인 협업필터링 추천시스템의 성능과 비교하였다.

추천정확도를 측정하기 하기 위한 지표로는 정확도(Precision)와 회상(Recall)이 가장 많이 알려져 있다(Herlocker, 2004). 정확도는 무비렌즈 전체 데이터셋에서 고객별로 추천한 영화들 중에 실제 고객이 좋다고 선호도를 표시한 영화의 비율을 나타내고, 회상은 고객별 선호도를 표시한 영화들 중에 추천시스템이 실제 추천한 영화의 비율을 나타낸다. 하지만 이 둘은 서로 trade-off 관계에 있으므로 최근에는 F 값을 많이 활용하고 있으며, 본 실험에서도 이 F 값을 최종 측정지표로 사용하였다 (Sarwar et al., 2000).

$$F = 2 \times \frac{Recall \times Precision}{Recall + Precision} \qquad (2)$$

4. 실험결과

4.1 원래의 협업필터링 추천시스템

데이터 튜닝을 하지 않은 원본 데이터셋을 일반 협업필터링 추천알고리즘을 사용하여 시뮬레이션하였다. 시뮬레이션을 위한 프로그램은 R을 사용하였다. 결과는 <Table 1>과 같이 이웃의 숫자(Neighbor size)가 70일 때, 추천 개수를 1, 3, 5개 F의 평균값이 0.0475로 최고의 성과가 나타났다. 이 값을 뒤의 SNA를 적용한 협업필터링 추천시스템의 성과와 비교하였다.

4.2 소셜 네트워크 분석 기법을 활용한 협업필터링 추천시스템

전체 943개의 노드들이 있는 데이터셋에서 연

결 정도 중앙성이 1600이하인 노드 41개를 분리하고 902개 노드들의 데이터셋에 대하여 일반 협업필터링 알고리즘으로 시뮬레이션 하였다. 이때 1600이라는 숫자는 시뮬레이션을 통해서 추천시스템의 정확도가 최대가 되는 숫자를 구한 결과이다. <Table 2>는 남겨진 902개의 노드로 이루어진 데이터셋에서 코사인 유사도 기법으로 이웃의 숫자가 70명일 때 최적의 추천성과를 내는 것을 보여주고 있다. 시뮬레이션을 위하여 랜덤샘플링 방식으로, 902개 노드 데이터셋 전체를 트레이닝셋 80%, 테스트셋 20% 비율로 나누어 실시하였다.

〈Table 1〉 F Measure of Whole Dataset (User-based CF with Cosine Measures)

NN	Number of Recommandation			Average
	1	3	5	
20	0.019	0.045	0.066	0.0434
30	0.019	0.048	0.071	0.0458
40	0.018	0.048	0.073	0.0462
50	0.019	0.049	0.071	0.0463
60	0.019	0.050	0.073	0.0473
70	0.019	0.050	0.074	0.0475

〈Table 2〉 F Measure of 902 Nodes (User-based CF with Cosine Measures)

NN	Number of Recommendation			Average
	1	3	5	
20	0.0201	0.0508	0.0758	0.0489
30	0.0209	0.0533	0.0800	0.0514
40	0.0211	0.0535	0.0798	0.0515
50	0.0211	0.0541	0.0798	0.0517
60	0.0216	0.0553	0.0798	0.0522
70	0.0214	0.0553	0.0808	0.0525

<Table 3>은 연결 정도 중앙성이 1600이하인 41개 노드들로 구성된 데이터셋에 대한 결과이다. 독특한 취향의 사용자들(gray sheep)이 몰려있는 이 데이터셋에서는 코사인 유사도 기법으로는 추천생성이 불가능하므로, 베스트셀러 추천기법으로 시뮬레이션을 진행하였다. 정확한 추천성능 비교를 위하여 트레이닝셋을 노드를 분리하기 전의 943개 노드가 모두 포함된 전체 데이터셋으로 지정하고, 테스트셋을 지표로 분리한, 41개 노드가 들어있는 데이터셋으로 하였다.

마지막으로 성능비교에 더욱 공정성을 기하기 위하여, <Table 4> 에서 보여지는 바와 같이 코사인 유사도를 적용하여 최적의 성과를 구한 902개 데이터셋과 베스트셀러 유사도를 적용하여 추천성능을 구한 41개 노드 데이터셋에 대하여 노드수들의 비율로 가중평균을 구하였다.

〈Table 3〉 F Measure of 41 Nodes (Popular Items)

NN	Number of Recommendation			Average
	1	3	5	
Popular	0.011	0.020	0.045	0.025

〈Table 4〉 Weighted Average F Measures of SNA + CF

NN	Number of Recommendation			Average
	1	3	5	
F	0.021	0.054	0.080	0.052

4.3 성과 비교

제시된 방법의 성과를 측정하기 위하여 본 기법을 적용하지 않은 추천시스템과 본 기법을 적

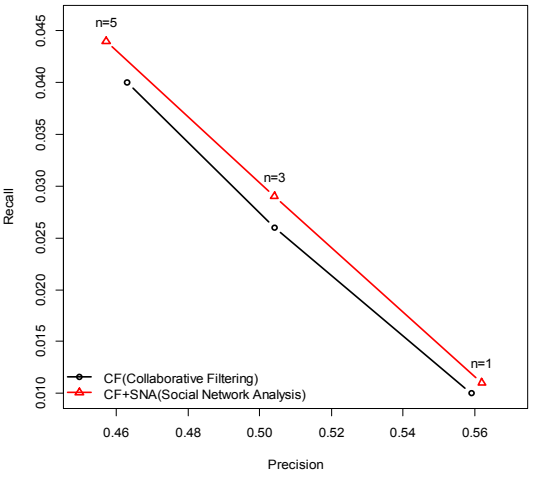
용한 추천시스템을 각각 최고의 성능을 내도록 시뮬레이션하여 성능을 비교하였다.

시뮬레이션 결과, <Table 5>와 같이 소셜 네트워크분석(SNA) 기법을 적용하지 않은 원래의 협업필터링(CF) 추천시스템 알고리즘으로는 F값이 0.0475였고, 소셜 네트워크분석(SNA) 기법을 적용한 협업필터링(CF) 추천시스템의 F값은 0.052로 추천성능이 평균 9.5% 향상되었음을 검증하였다. <Figure 1> 의 그래프를 통하여 본 연구에서 개발한 알고리즘에 의해 회상(Recall) 값이 크게 개선되는 것을 볼 수 있다. 이는 추천생성을 위해 고객들간의 선호 아이템의 유사성을 계산할 때, 다른 고객들과 유사성이 적은 독특한 고객들(gray sheep)을 분리하여 따로 계산하였기 때문에 나머지 대부분 고객들의 유사성을 계산함에 있어 정확도가 향상된 결과라고 생각된다. <Figure 2>의 그래프에서와 같이 추천 개수가 늘어날수록 성능향상의 차이가 더욱 커지는 것을 확인할 수 있다.

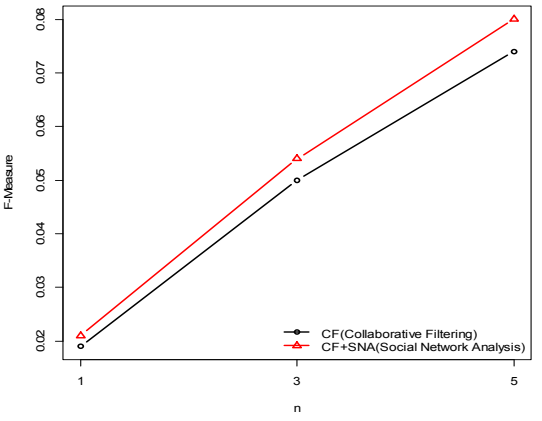
추천갯수가 1, 3, 5 개 일 때, 각각의 F값을 구하여 비교하면 추천갯수가 1(n=1)일 경우에는 최대 11%까지 성능이 향상되었다.

〈Table 5〉 Comparison of Performances

Top N List	Number of Recommendation			Average
	N=1	N=3	N=5	
Original CF (F measure)	0.019	0.05	0.074	0.0475
SNA + CF (F measure)	0.021	0.054	0.08	0.052
Improvements	11%	8%	8%	9%



〈Figure 1〉 Precision & Recall



〈Figure 2〉 F Measures

5. 결론

본 연구에서는 소셜 네트워크 분석(Social Network Analysis: SNA) 기법을 활용하여 협업필터링 (Collaborative Filtering: CF) 추천시스템의 성능을 향상시키고자 데이터 튜닝을 통한 특이한 취향의 고객들을 분리하는 방법을 개발하였

고 시뮬레이션을 통해 이를 검증하였다. 검증 결과, 가중평균을 적용한 개발모델이 일반적인 협업필터링의 모델보다 평균 10%의 성능향상을 보였다.

그러나 무비렌즈 데이터 한 종류로 모형을 검증을 하였기 때문에 향후 좀 더 다양한 도메인의 데이터에서 검증이 필요하다. 또한 본 연구에서 제안하는 모형의 절대적인 성과가 낮다는 한계점이 존재한다. 추후 추천 성능을 개선하기 위하여 추가적인 알고리즘과 데이터 셋팅의 결합에 대한 연구가 이루어져야 할 필요가 있다.

그럼에도 불구하고 본 연구는 이론적인 측면에서 SNA 방법을 적용해서 CF의 정확도 향상의 여지를 보여줬으며 앞으로 다른 측면에서의 정확도 향상의 밑거름이 될 수 있을 것으로 기대된다. 실무적으로는 이 연구의 성과는 추후 전자상거래 기업들이 추천 시스템을 구축하고 고객들에게 맞춤형 개인화 서비스를 실시함에 있어, 비교적 간단하게 적용하여 추천성능을 향상시킬 수 있는 알고리즘이 될 것으로 기대된다.

참고문헌 (References)

Ahn, S. M., I. H. Kim, B. G. Choi, Y. H. Cho, E. H. Kim, and M. Y. Kim, “Understanding the Performance of Collaborative Filtering Recommendation through Social Network Analysis” *Society for e-Business Studies*, Vol.17, No.2(2012), 129-147.

Barragans-Martinez, A. B., E. Costa-Montenegro, J. C. Burguillo, M. Rey-Lopez, F. A. Mikic-Fonte and A. Peleteiro, "A Hybrid Content-Based and Item-Based Collaborative Filtering Approach to Recommend Tv Programs Enhanced with Singular Value Decomposition," *Information Sciences,* Vol.180, No.22(2010), 4290-4311.

Claypool, M., A. Gokhale, T. Miranda, P. Murnikov, D. Netes, and M. Sartin. "Combining Content-Based and Collaborative Filters in an Online Newspaper," *Proceedings of ACM SIGIR workshop on recommender systems*, (1999).

Dorogovtsev, S. N. and J. F. F. Mendes, "Evolution of Networks," *Advances in Physics,* Vol.51, No.4 (2002), 1079-1187.

Ghazanfar, M. A. and A. Prügel-Bennett, "Leveraging Clustering Approaches to Solve the Gray-Sheep Users Problem in Recommender Systems," *Expert Systems with Applications,* Vol.41, No.7 (2014): 3261-3275.

Goldberg, D., D. Nichols, B. M. Oki, and D. Terry, "Using Collaborative Filtering to Weave an Information Tapestry," *Communications of the Acm,* Vol.35, No.12(1992), 61-70.

Herlocker, J. L., J. A. Konstan, and J. T. Riedl, "An Empirical Analysis of Design Choices in Neighborhood-Based Collaborative Filtering Algorithms," *Information Retrieval,* Vol.5, No.4 (2002), 287-310.

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

Hung, L. P., "A Personalized Recommendation System Based on Product Taxonomy for One-to-One Marketing Online," *Expert Systems with Applications,* Vol.29, No.2(2005), 383-392.

Im, I. and A. Hars, "Does a One-Size Recommendation

System Fit All? The Effectiveness of Collaborative Filtering Based Recommendation Systems across Different Domains and Search Modes," *ACM Transactions on Information Systems (TOIS),* Vol.26, No.1(2007).

Kim, H. K., J. K. Kim, and Q.-Y. Chen, "A Network Approach to Derive Product Relations and Analyze Topological Characteristics", *Journal of Intelligence and Information Systems,* Vol.15, No.4 (2009), 159-182.

Kim, Y. H., *Social Network Analysis*, Pakyoungsa, Seoul, 2013.

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

Konstan, J. A., J. Riedl, A. Borchers, and J. L. Herlocker, "Recommender Systems: A Grouplens Perspective," In *Recommender Systems: Papers from the 1998 Workshop (AAAI Technical Report WS-98-08)*, (1998), 60-64.

Lee, Y. J., S. H. Lee,and C. J. Wang, "Improving Sparsity Problem of Collaborative Filtering in Educational Contents Recommendation System,", *Proceedings of Korea Information Science Society,* Vol.30, No.1(A)(2003), 830-832.

Newman, M. E. J., "The Structure and Function of Complex Networks," *Siam Review,* Vol.45, No.2 (2003), 167-256.

Park, J. H. and Y. H. Cho, "Social Network Analysis for the Effective Adoption of Recommender Systems," *Journal of Intelligence and Information Systems,* Vol.17, No.4(2011), 305-316.

Pham, M. C., Y. Cao, R. Klamma, and M. Jarke, "A Clustering Approach for Collaborative Filtering Recommendation Using Social Network Analysis," *Journal of Universal Computer Science,* Vol.17, No.4(2011), 583-604.

Sarwar, B., G. Karypis, J. Konstan, and J. Riedl, "Analysis of Recommendation Algorithms for E-Commerce," *Proceedings of the 2nd ACM conference on Electronic commerce*, (2000), 158-167.

Shin, C. H., J. W. Lee, H. N. Yang, and I. Y. Choi, "The Research on Recommender for New Customers Using Collaborative Filtering and Social Network Analysis,", *Journal of Intelligence and Information Systems,* Vol.18, No.4(2012), 19-42.

Sohn, D. W., *Social Network Analysis*, Kyungmoon publishers, Seoul, 2008.

Su, X. and T. M. Khoshgoftaar, "Collaborative Filtering for Multi-Class Data Using Bayesian Networks," *International Journal on Artificial Intelligence Tools,* Vol.17, No.1(2008), 71-85.

Abstract

Resolving the 'Gray sheep' Problem Using Social Network Analysis (SNA) in Collaborative Filtering (CF) Recommender Systems

Minsung Kim · Il Im*

Recommender system has become one of the most important technologies in e-commerce in these days. The ultimate reason to shop online, for many consumers, is to reduce the efforts for information search and purchase. Recommender system is a key technology to serve these needs. Many of the past studies about recommender systems have been devoted to developing and improving recommendation algorithms and collaborative filtering (CF) is known to be the most successful one. Despite its success, however, CF has several shortcomings such as cold-start, sparsity, gray sheep problems. In order to be able to generate recommendations, ordinary CF algorithms require evaluations or preference information directly from users. For new users who do not have any evaluations or preference information, therefore, CF cannot come up with recommendations (Cold-star problem). As the numbers of products and customers increase, the scale of the data increases exponentially and most of the data cells are empty. This sparse dataset makes computation for recommendation extremely hard (Sparsity problem). Since CF is based on the assumption that there are groups of users sharing common preferences or tastes, CF becomes inaccurate if there are many users with rare and unique tastes (Gray sheep problem).

This study proposes a new algorithm that utilizes Social Network Analysis (SNA) techniques to resolve the gray sheep problem. We utilize 'degree centrality' in SNA to identify users with unique preferences (gray sheep). Degree centrality in SNA refers to the number of direct links to and from a node. In a network of users who are connected through common preferences or tastes, those with unique tastes have fewer links to other users (nodes) and they are isolated from other users. Therefore, gray sheep can be identified by calculating degree centrality of each node. We divide the dataset into two, gray sheep and others, based on the degree centrality of the users. Then, different similarity measures and recommendation methods are applied to these two datasets. More detail algorithm is as follows:

* Corresponding Author: Il Im
School of Business, Yonsei University
50 Yonsei-ro, Seodaemun-gu, Seoul 120-749, Korea
Tel: +82-2-2123-5480, Fax: +82-2-2123-8639, E-mail: il.im@yonsei.ac.kr

Step 1: Convert the initial data which is a two-mode network (user to item) into an one-mode network (user to user).

Step 2: Calculate degree centrality of each node and separate those nodes having degree centrality values lower than the pre-set threshold. The threshold value is determined by simulations such that the accuracy of CF for the remaining dataset is maximized.

Step 3: Ordinary CF algorithm is applied to the remaining dataset.

Step 4: Since the separated dataset consist of users with unique tastes, an ordinary CF algorithm cannot generate recommendations for them. A 'popular item' method is used to generate recommendations for these users. The F measures of the two datasets are weighted by the numbers of nodes and summed to be used as the final performance metric.

In order to test performance improvement by this new algorithm, an empirical study was conducted using a publically available dataset – the MovieLens data by GroupLens research team. We used 100,000 evaluations by 943 users on 1,682 movies. The proposed algorithm was compared with an ordinary CF algorithm utilizing 'Best-N-neighbors' and 'Cosine' similarity method. The empirical results show that F measure was improved about 11% on average when the proposed algorithm was used <Table 8>.

Past studies to improve CF performance typically used additional information other than users' evaluations such as demographic data. Some studies applied SNA techniques as a new similarity metric. This study is novel in that it used SNA to separate dataset. This study shows that performance of CF can be improved, without any additional information, when SNA techniques are used as proposed.

This study has several theoretical and practical implications. This study empirically shows that the characteristics of dataset can affect the performance of CF recommender systems. This helps researchers understand factors affecting performance of CF. This study also opens a door for future studies in the area of applying SNA to CF to analyze characteristics of dataset. In practice, this study provides guidelines to improve performance of CF recommender systems with a simple modification.

Key Words : Collaborative filtering (CF), Social Network Analysis (SNA), Gray Sheep Problem

Received : June 15, 2014 Revised: June 19, 2014 Accepted: June 22, 2014

저 자 소 개

김 민 성
동덕여자대학교 컴퓨터학과 학사를 취득하고 엘지전자 MC연구소, 한국개발연구원(KDI) 경제개발협력연구실에서 근무하였으며 현재 연세대학교 경영학과 석사과정에 재학 중이다. 주요 관심분야는 추천시스템, 소셜 네트워크분석 응용, 데이터마이닝 이다.

임　일
서울대학교에서 경영학 학사와 석사를 취득하였으며, 미국 University of Southern California에서 경영학 박사를 받았다. 미국 New Jersey Institute of Technology의 Department of Information Systems에서 교수를 하였고 현재 연세대학교 경영학과 교수로 재직 중이다. 주요 연구분야는 추천시스템, SNS, 기술 수용 등이다.