物流學碩士　學位論文

사회연결망을 활용한 신규고객 추천 및 비추천시스템의 비교분석

Comparative Analysis of Recommendation and Nonrecommendation Systems for New Customers Using Social Networks.

指導敎授 辛　昌　勳

2012年　2月

韓國海洋大學校 大學院

物流시스템學科

양 한 나

物流學碩士　學位論文

사회연결망을 활용한 신규고객 추천 및 비추천시스템의 비교분석

Comparative Analysis of Recommendation and Nonrecommendation Systems for New Customers Using Social Networks.

指導敎授 辛　昌　勳

2012年　2月

韓國海洋大學校 大學院

物流시스템學科

양 한 나

- 목 차 -

- 표 목 차 -

- 그 림 목 차 -

Comparative Analysis of Recommendation and Nonrecommendation Systems for Nes Customers Using Social Networks.

Han Na, Yang

Department of Logistics Engineering Graduate School of

Korea Maritime University

Abstract

With the rapid growth of internet and the proliferation of electronic commerce, the frequency of the customers buying the products through internet is also increasing. But, since there is a big variety of goods, it is not easy for a customer to find the product that suits his intention of purchase. That the corporation recommend the products that suits the purchase intention of the customers is very important since it can give satisfaction to the customer which can in turn lead to the magnification of the sales. According to this, a recommendation system has started to getting attention. There are various methods in the recommendation system. But, for the new customers who have not the purchasing history, the recommendation is difficult since we can not predict their preferences.

Accordingly, in this thesis, we present the recommendation system and non-recommendation system for new customers by using the cooperative filtering and Social Network Analysis(SNA). In order to assess the

proposed systems, we have applied the data of MovieLens and analysed the result.

제 1 장 서 론

제 1 절 연구의 배경 및 목적

인터넷의 급속한 성장과 더불어 전자상거래가 확산되면서 고객이 인터넷을 통해 상품을 구매하는 빈도도 증가하고 있다. 인터넷을 이용하여 상품을 구매하는 고객은 편리성과 신뢰성을 기본으로 하고 있다. 그러나 상품의 폭과 깊이가 점점 더 다양해지고, 많은 상품들이 쏟아져 나오는 현실 속에서 자신의 구매의도와 완벽하게 일치하는 상품을 구매하는 것은 점점 더 어려워지고 있다. 인터넷을 이용하여 상품을 판매하는 기업의 입장에서도, 고객의 구매의도와 완벽하게 일치하는 상품을 가시적으로 보여주는 것은 매출에 큰 영향을 미치는 만큼 그 중요성이 더욱 커지고 있다. 이에 따라 CRM(Customer Relationship Management, 고객관계관리) 등의 방법을 통해 고객의 니즈(needs)를 충족시키기 위한 노력이 확대되면서 추천시스템(Recommender System)이 주목받기 시작했다. 추천시스템은 고객이 원하는 상품 혹은 서비스를 정확히 예측하여 추천하는 시스템이다. 따라서 고객 정보가 많으면 많을수록 그 정확도가 더욱 높아지는 것은 당연한 것이다. 그러나 취급하는 상품과 고객의 데이터가 증가할수록 완벽한 데이터의 수집 및 추천은 어려워진다. 이처럼 데이터 처리 문제의 해결과 동시에 정확도도 증가시키기 위한 수많은 추천방법들이 개발되어 왔다.

추천시스템은 인구통계정보기반 추천방법, 베스트셀러기반 추천방법, 최소질의대상 상품결정방법, 내용필터링, 협력필터링 등이 있으며 추천 정확성을 볼 때, 협력필터링 방식이 다른 방식보다 비교적 정확도가 높은 것으로 알려져 있다. 이는 과거에 구매를 했던 데이터가 존재하는 기존고객과 유사한 구매형태 혹은 패턴을 보이는 고객의 구매정보를 이용하여 선호도가 높을 것으로 예측되는 상품을 추천해 주는 방식이다. 그러나 과거 구매 이력이 전혀 없는 신규고객의 경우 선호도를 예측할 수 없으므로 추천이 어렵게 된다.

이와 대조적으로 베스트셀러기반 추천방법은 많이 팔린 순서로 상품을 추천하여 기존고객 뿐 아니라 신규고객에게도 추천이 가능하다. 그러나 불특정다수인 고객의 선호도이므로 모든 고객에게 맞춤화된 정보를 제공하는 것에는 어려움이

있다. 따라서 신규고객에게 상품을 추천하는 기존 방식의 한계를 극복할 수 있는 새로운 추천방법이 필요하다.

최근 사회연결망분석(Social Network Analysis)에 관한 연구가 활발히 진행되고 있다. 사람과 사람 혹은 집단과 집단의 관계를 분석하고 각 개체 간 존재하는 연결구조에 대한 가시적인 결과도 볼 수 있는 분석기법이다.

본 논문에서는 협력 필터링과 사회연결망 분석의 중심성을 이용한 추천시스템 및 비추천시스템을 제시한다. 기존 추천시스템의 단점을 보완하기 위하여 고객 데이터 수집 비중은 낮추고, 정확성은 높일 수 있는 분석기법을 제시한다. 중심성이 높은 기존고객들의 정보를 이용한 추천시스템과 대부분의 기존 연구에서 사용되지 않았던 중심성이 낮은 고객들의 정보를 활용하여 비추천시스템을 살펴보고 새로운 시사점을 도출하고자 한다.

제 2 절 연구의 내용 및 구성

2.1 연구의 내용

인터넷을 이용한 상품의 판매는 고객의 니즈를 얼마나 정확하게 파악하고 추천해 주느냐에 따라 매출이 크게 좌우된다. 클릭 몇 번만으로 한 개인의 구매 욕구를 해결할 수 있을 만큼 편리해진 현 시점에서, 인터넷을 통한 구매에 대한 고객의 의존도와 활용도는 더욱 높아지고 있다. 이에 따라 상품추천방식에 관한 연구도 함께 증가하고 있다. 그러나 고객의 과거 구매이력 및 고객관련 데이터가 부족한 신규고객의 경우 선호하는 상품이나 구매 패턴을 파악하는 것은 쉽지 않다.

최근 기존고객에게 상품을 추천하기 위한방법으로 협력필터링이 많이 사용되고 있지만 이 방법도 기존 구매이력이 없는 신규고객에게 적용하기에 문제가 있다. 베스트셀러기반 추천방법도 추천과정이 간단하며 많은 데이터의 저장을 요구하지 않는다는 장점이 있지만 모든 고객에게 동일한 상품을 추천하게 되어 개인화된 추천을 하지 못하며, 추천된 상품이 특정 상품 군에 한정적인 경우가 많아 정확도가 낮은 것으로 알려져 있다. 따라서 이러한 기존 추천 방법들의 한계를 극복할 수 있는 새로운 연구가 필요하다.

최근 사회연결망 분석의 중심성(centrality)을 활용한 연구가 활발히 진행되고 있다. 기존고객들의 상품구매 패턴을 분석하고 유사도가 높은 고객들 간의 관계망을 형성한 후, 중심성이 높은 고객들을 찾고 추천 그룹을 형성하여 그들이 구매한 상품을 신규고객에게 추천하는 방식이다.

중심성이란 전체 네트워크에서 한 사람의 위치가 가지는 중심의 정도를 표현하는 지표이다(손동원, 2002). 네트워크에서의 중심은 가시성이 높을 경우, 다른 사람과 관계를 가지는 정도가 높은 경우, 중개자의 위치인 경우 등의 다양한 개념으로 해석된다.

기존고객 집단 내에서 중심성이 높은 고객은 신규고객과도 상품의 구매 유사도가 일치할 가능성이 높으며, 구매 유사도가 일치할 경우 추천 정확도도 더욱 높아질 것이라고 가정한다. 또한 중심성이 낮은 고객들의 연결망을 통하여 생성된

구매 목록은 다시 말하면 신규고객에게 비추천 해야 할 목록이라고 가정한다.

본 논문에서는 중심성이 높은 고객들의 pls(구매가능점수:purchase likelihood score) 가운데 상위 목록을 추천 할 목록, 그리고 중심성이 낮은 고객들의 pls 가운데 상위 목록을 비추천 할 목록으로 두 그룹으로 나누어 분석한다. 그 과정에서 생성된 각각의 목록을 신규고객에게 추천하고, 그 적중률을 비교한다.

2.2 연구의 구성

본 연구는 구성은 다음과 같다. 1장에서는 연구의 배경 및 목적과 구성을 제시하였고, 2장에서는 기존의 추천기법에 관한 정의 및 한계점, 그리고 사회 연결망 분석에 관한 선행연구를 실시하였다. 그리고 3장에서는 본 연구를 위한 방법론을 설명하고 4장에서 분석 및 결과의 도출 후, 5장에서 연구결과 요약과 시사점 및 한계에 관하여 제시한다.

제 2 장 선행연구

제 1 절 추천시스템(Recommender System)

인터넷에서의 상품추천시스템은 고객과 상품에 관한 정보를 분석하여 고객이 관심을 가질만한 상품을 찾아내고 그 상품을 고객에게 추천해주는 것이다. 최근에는 인구통계정보기반 추천방법, 베스트셀러기반 추천방법, 정보필터링기법, 최소질의대상 상품 결정방법 등 다양한 추천방법을 통한 연구가 진행되고 있다.

먼저 인구통계정보기반 추천방법은 고객이 인터넷 쇼핑몰에 가입할 때 입력한 인구통계정보를 이용하는 방법이다. 즉, 신규고객이 인터넷 쇼핑몰에 가입 시, 자신의 정보를 입력하면 고객이 입력한 정보를 분석하여 고객에게 추천 상품을 가시적으로 보여주는 시스템이다. 그러나 고객이 정보를 입력할 때 느낄 수 있는 번거로움으로 인하여 성실하지 못한 답변을 하거나 누락될 가능성이 있으므로 신뢰도가 낮다. 또한 인구통계정보와 구매 간의 상호 관련성이 높지 않으므로 추천 정확도가 높지 않은 것으로 보고되고 있다.

베스트셀러기반 추천방법은 협력필터링 방식을 보완하기 위한 대안(Sarwar et al. 2000)으로 가장 많이 팔린 순서로 상품을 추천하는 방법이다. 신규고객이 쇼핑몰에 접속하면 고객의 정보 없이도 단시간 내에 상품을 추천해 줄 수 있다는 장점이 있어 실무에서 많이 활용되고 있다. 그러나 각 고객에게 맞춤화되고 개인화된 추천을 할 수 없으며, 많이 팔린 상품들이 특정 상품 군에 속하는 경우가 많기 때문에 추천 정확도가 낮은 것으로 알려져 있다(박종학 외. 2009).

정보필터링기법은 내용필터링과 협력필터링(또는 협업필터링)으로 나눌 수 있다(Lee et al. 2001; Kohrs and Merialdo, 2001). 정보필터링기법은 타 기법에 비해 고객에게 제공되는 정보의 정확도가 높다는 장점 때문에 많이 활용되고 있다. 신규고객의 경우 과거 구매이력이 존재하지 않기 때문에 고객의 구매정보를 활용한 추천이 불가능하고, 이웃고객의 형성 자체가 불가능하다는 문제점이 있다.

정보필터링기법의 중류 중 하나인 내용필터링은 해당 고객이 과거에 선호했던 상품과 가장 유사한 상품을 고객에게 추천하는 것이고, 협력필터링은 해당 고객과 가장 유사한 선호도를 가진 이웃고객을 찾아내고 이들 이웃고객의 선호정보를 기반으로 하여 상품을 추천하는 것이다. 협력필터링은 이웃고객의 평가를 수집하는 방법을 통하여 해당 고객의 선호도나 정보를 찾는 시간을 줄일 수 있다는 장점이 있다(이용준 외. 2003). 그러나 Sarwar et al. (2000)은 가장 성공적인 추천기법 중 하나가 협력필터링이라고 주장하였지만, 데이터의 희박성 문제, 알고리즘의 확장성 문제와 추천의 질적 문제와 같은 단점도 존재한다고 했다.

최소질의대상 상품 결정방법은 신규고객에게 상품목록을 제시하고 그에 관한 선호도를 고객이 직접 응답하도록 하는 것이다. 이 방법은 선호도 입력이라는 간단한 방법으로 고객에 대한 정보 파악이 가능한 방법이다. 이 때 질의되는 상품목록을 선정하는 방법에 관한 연구도 다양하게 진행되고 있다. 그러나 고객들이 직접 선호도를 입력해야 하는 번거로움이 단점으로 지적되고 있다. 또한 선호도 입력 시, 불완전한 답변을 하게 될 경우 정보의 정확성이 저하되며 이에 따라 추천의 정확성 역시 저하된다는 단점이 존재한다(Schein et al. 2002; Yu et al. 2004; 박종학 외. 2009).

이처럼 추천시스템에 관한 연구들이 계속 연구되고 있다. 이용준 외. (2003)의 연구에서는 협업 여과 추천계산의 희소성으로 발생되는 추천의 정확도를 보완하기 위해 인구통계정보를 이용하여 가상평가점수를 부가하고, 유사도 계산의 정확도를 향상시켜 예측의 정확도를 높이는 방식을 제안했다. 신동원(2006)은 협업필터링 방식과 사회연결망 분석을 활용하여 신규고객에게 상품을 추천하는 방식을 제안했다. 기존고객 중 사회 연결망분석에서 중심성이 높은 고객이 향후 신규고객에게 추천할 확률이 높기 때문에 중심성이 높은 기존고객의 구매상품을 추천하는 방법을 제시하였다. 박종학 외. (2009)는 협력필터링에 사회연결망의 중심성 분석기법을 적용하여 신규고객의 잠재적인 이웃고객을 찾고, 그들의 구매정보를 이용하여 신규고객에게 상품을 추천하는 방법을 제시하였다. 강부식(2010)의 연구에서는 협력필터링과 사회연결망의 구조적 공백 개념을 활용한 방식을 제안하고 베스트셀러기반 방식과 추천 정확성을 비교하였다. 이재식과 박석두(2007)는 장르별 협업필터링 방법을 통하여 희박성과 확장성

문제를 극복하는 방법을 제안하였는데 영화 데이터를 사용하여 GenreWise_CF를 개발하여 장르추천 적중률을 향상시켰다.

고객을 위한 다양한 추천기법이 연구되고 있지만 본 논문에서는 가장 성공적인 상품추천방법으로 알려져 있는 협력필터링 방법을 사용한다. 협력필터링이 가지고 있는 신규고객 추천에 대한 단점을 보완하기 위하여 사회연결망의 중심성을 이용하여 선호도를 예측한다. 또한 기존 연구에서는 대부분 사용되지 않고 있는 중심성이 낮은 고객들의 선호도 데이터를 활용하여 신규고객에게 추가적인 정보를 제공하는 방법인 비추천시스템을 통하여 새로운 시사점을 제시하는데 그 목적이 있다.

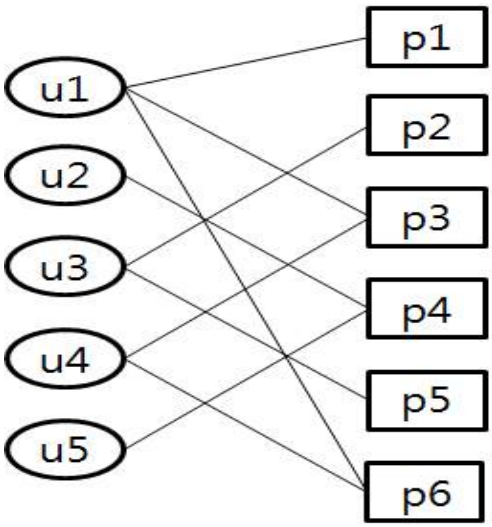
제 2 절 사회연결망 분석(Social Network Analysis)

사회연결망은 Barnes(1954)에 의해 처음 사용된 용어로 개인적인 인간관계가 확산되어 형성된 사람들 사이의 연결된 네트워크이다(손동원, 2002). 관계적 인간관에 입각하여 인간 행위와 사회구조의 효과를 설명하려는 시도(김용학, 2003)로, 관계내의 상호작용 및 특성을 찾는데 그 목적이 있으며 시각적인 효과 표현도 가능하다.

배순한 외. (2010)은 시간의 변화에 따른 커뮤니티의 중심성 변화를 연구하였다. 사회연결망 분석방법 가운데 하나인 중심성을 통해 온라인에서의 각 개인의 영향력에 대하여 분석하였는데 이 연구는 하나의 온라인 커뮤니티에 관한 것으로 일반화시키기에는 부족하다는 단점이 있었다. 국승용(2007)은 농산물 물류센터의 적정입지 선정을 위한 환적모형 활용 결과에서 생성된 행렬을 NetMiner Professional Edition Version 2.6을 사용하여 연결망을 분석하였다. 취급 비중이 높은 물류센터가 중심성이 높은 것을 보여주었는데 사회연결망 분석을 보완적으로 활용하여 더욱 풍부한 정보를 나타내기 위한 시도를 하였다. 이기헌과 김창욱(2010)은 신제품의 확산 과정을 재 연결 확률을 이용하여 작은 세상 연결망을 활용한 연구를 하였으며, 이지선과 강신겸(2010)은 증도를 대상지로 선정하여 이해관계자간 협력관계에 대하여 연구 하였다. 커뮤니티관광개발 계획 초기단계에서 나타나는 관계자들 간 협력관계의 특징을 분석하였다. 이처럼 사회연결망은 다양한 분야에서 응용되고 있다.

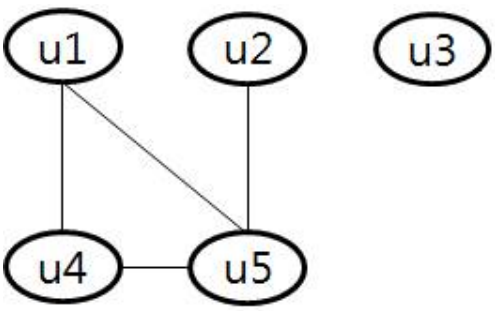
연결망은 점과 연결선으로 만들어지고, 네트워크 내에서 n개의 점은 n*n 형태의 행렬로 나타낸다. 이때 각 개체 간 연결이 존재하면 1, 존재하지 않으면 0으로 표현한다. 연결망 분석을 위해서는 분석 단위와 그에 적합한 자료의 형태가 정해져야 하는데 직접적인 상호작용의 관계가 없더라도 관계를 인위적으로 설정한 연결망을 준 연결망(quasi network)이라고 한다(김용학, 2003).

아래 < 그림 2-1 >은 준 연결망의 예로 고객과 해당 고객이 구매한 상품과의 연결을 나타낸다(박종학 외. 2009).



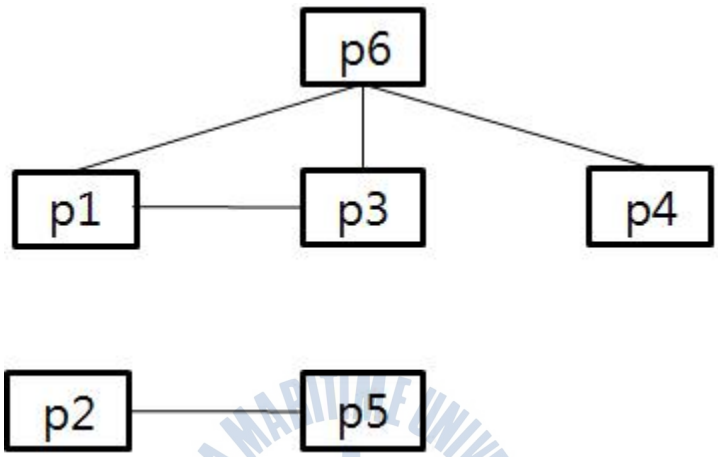
< 그림 2-1 > 구매 트랜잭션

< 그림 2-2 >는 위 < 그림 2-1 >에서 연결한 고객과 상품간의 네트워크를 고객 간의 관계로 표현한 것이다. 동일 제품을 구매한 고객과 고객은 서로 직접적 상호작용의 관계라는 것을 표현하는 연결망은 < 그림 2-2 >와 같은 형태로 나타난다.



< 그림 2-2 > 고객 네트워크

아래의 < 그림 2-3 >은 제품 네트워크로 제품과 제품 간의 관계를 나타낸 그림이다. 한 고객이 구매한 제품들의 관계에도 역시 상호관계가 있다는 것을 연결망의 형태로 나타낸 것이다.



< 그림 2-3 > 제품 네트워크

사회연결망 분석방법 가운데 연결망의 구조를 파악하고 특성을 살펴보기 위한 기법으로 밀도(density), 구조적 공백(또는 구조적 틈새:structural hole), 중심성(centrality) 등이 있다(Bonacich, 1987; 손동원, 2002; 김용학, 2003).

밀도란 한 연결망에서 행위자들 사이의 연결된 정도를 의미하는 것으로 전체 구성원이 얼마나 많은 관계를 맺고 있는가를 표현하기 위한 개념이다. 한 연결망이 얼마나 완벽한가를 표현하는 것으로 완벽한 밀도는 모든 점들이 서로 연결되어 있는 상태에서 성립한다(손동원, 2002).

구조적 공백은 미국의 사회학자인 Ronald Burt에 의해서 주장된 개념으로 한 사람이 다른 사람들과의 연계에서 중복되지 않고 그 행위자를 통해서만 연계되는 위치를 의미한다(손동원, 2002). 중복적인 연결을 피해 다양한 그룹들과 연결될 때 효율을 증가시킬 수 있다(Burt, 1992).

중심성은 한 행위자가 전체 연결망에서 중심에 위치하는 정도를 표현하는 지표로 연결정도중심성, 근접중심성, 매개중심성 등을 통하여 측정할 수 있다. 연결정도(dagree) 중심성은 다른 점과 연결된 정도를 중심으로 보는 개념으로 연결된 점의 수가 많거나 적음의 여부가 기준이 되며, 근접(closeness)중심성은 한 점이 다른 점에 얼마만큼 가까운가에 관한 개념으로 두 점 사이의 거리가 기준이 된다. 매개(betweenness)중심성은 한 점이 얼마나 다른 점들과의 연결망을 구축하는데 중개자 역할을 하느냐를 말하는 개념으로 중개역할에 초점을 둔다. (손동원, 2002).

구조적 공백은 한 고객이 얼마나 다양한 고객들과 비 중복적인 관계를 맺는 것이 가능한지와 관련된 개념이라면 중심성은 연결망 내에서 한 고객이 얼마나 많은 고객과 관계를 맺고 있는가를 측정하는 척도이다(강부식, 2010).

제 3 장 방법론

협력필터링의 문제점인 신규고객 상품추천에 대한 문제를 해결하기 위하여 사회연결망의 중심성을 활용하여 추천하고자 한다. 다음 < 표 3-1 >은 신규고객 추천방법의 절차를 나타낸다.

첫째, 기존고객의 구매유사도 분석 단계이다. 고객과 고객의 구매상품이 얼마나 일치하는지를 분석하는 단계이다.

둘째, 기존고객의 네트워크 구성 단계이다. 구매유사도를 이용해 일정 값 이상을 1로 정의하여 고객들을 링크로 연결한다.

셋째, 신규고객의 추천 이웃고객 형성 단계이다. 고객들의 중심성을 구하고 값이 높은 고객을 이웃고객으로 지정한다.

넷째, 신규고객 상품 추천 단계이다. 중심성이 높은 고객들의 구매상품을 찾아 상품의 구매가능성을 계산하여 구매가능성이 높은 상품을 신규고객에게 추천한다.

< 표 3-1 > 신규고객 추천방법의 절차

제 1 절 기존고객의 구매유사도 분석

신규고객에게 상품을 추천하는 첫 번째 단계는 기존고객의 구매유사도 분석이다. 기존고객들의 구매성향을 분석하여 고객들의 유사도를 분석하는 단계로 현재 데이터 내의 기존고객이 어떤 상품을 구매했는지를 파악하는 것이다. 식 (1)과 같이 고객-구매상품 매트릭스로 표현한다. 상품을 구매했을 경우 1, 구매하지 않았을 경우 0으로 표현하며, 같은 상품을 한 고객이 여러 번 구매하였더라도 해당 상품은 한 가지 이므로 1로 표현한다.

$$P_{ij} = \begin{cases} 1 : \text{고객 } i\text{가 상품을 구매} \\ 0 : \text{고객 } i\text{가 상품을 비구매} \end{cases} \quad (1)$$

매트릭스가 완성되면, 협력필터링을 통한 상품추천 방법을 위한 유사도를 계산한다. 식 (2)와 같이 피어슨 상관계수를 이용하여 계산하는데 고객들 사이의 유사도는 -1에서 +1 사이의 값을 갖는다. 유사도 값이 1이라면 두 고객의 구매 상품이 완벽히 일치하는 것이고 -1은 두 고객의 구매 상품이 완벽히 다르다는 것을 의미하며, 0이면 선형적인 측면에서 구매 유사도가 없다는 것을 의미한다. p_{ak}와 p_{bk}는 상품 k에 대한 고객 a와 b의 구매여부를 나타내며 $\overline{p_a}$와 $\overline{p_b}$는 고객 a와 b의 평균선호도를 의미한다.

$$s(a,b) = corr(a,b) = \frac{\sum_{k=1}^{L} (p_{ak} - \overline{p_a})\ (p_{bk} - \overline{p_b})}{\sqrt{\sum_{k=1}^{L} (p_{ak} - \overline{p_b})^2 \ \sum_{k=1}^{L} (p_{bk} - \overline{p_b})^2}} \quad (2)$$

제 2 절 기존고객의 네트워크 구성

신규고객에게 상품 추천을 위한 두 번째 단계는 앞서 계산한 유사도를 이용하여 유사한 구매패턴을 보이는 기존고객간의 연결망을 구성하는 것이다. 연결망은 점[1], 선[2] 등으로 이루어진다. 고객 간 구매유사도가 임계치 ρ 이상인 경우 링크는 1이 되고, 그렇지 않은 경우 0이 된다. 1은 고객(노드)간 연결이 있음을 의미하며 0은 연결이 없음을 의미한다.

$$link(a,b) = \begin{cases} s(a,b) \geq \rho : 1\,(\text{연결됨}) \\ s(a,b) < \rho : 0\,(\text{연결되지 않음}) \end{cases} \quad (3)$$

위의 수식 (3)은 고객 a와 b의 연결 link(a,b)를 의미한다. 연결된 링크가 많은 고객은 이웃이 많다는 것, 즉 유사한 구매패턴을 보이는 이웃고객이 많음을 의미한다(박종학 외. 2009).

1) 기존 연구에서는 node, vertex, point 등의 다양한 표현이 사용되고 있음
2) 기존 연구에서는 edge, line 등의 다양한 표현이 사용되고 있음

제 3 절 신규고객의 추천 이웃고객 형성

세 번째 단계는 신규고객의 이웃고객을 형성하는 단계이다. 사회연결망의 중심성을 구하고 중심성이 높은 고객을 선택하는데, 이때 중심성이 높은 고객이 이웃고객이 된다. 다시 말하면 신규고객의 상품선택에 가장 크게 영향을 미칠 것이라고 예상되는 고객이다. 중심성에서 값이 높게 나온 상위 K명을 뽑아 신규고객의 이웃고객 H로 가정하고 이웃고객이 구매한 상품을 신규고객에게 추천한다. 중심성이 높은 이웃고객은 다른 고객들에 비해 고객 간의 유사도가 높은 것으로, 구매 상품의 수와 관계없이 얼마나 다른 고객들과 유사한 상품을 구매했느냐에 따라 결정된다.

$$cen(c) = \sum_{j=1}^{M} link(c,j) \tag{4}$$

식 (4)에서 고객 c의 중심성 cen(c)은 고객 c에 직접 연결되어 있는 노드의 수를 구하는 것이다(박종학 외. 2009).

제 4 절 신규고객 상품 추천

앞서 선정한 중심성이 높은 고객(K명)들의 과거구매 상품 목록을 찾아 각각의 구매가능성을 계산하여 점수가 높은 상품(상위 N개)을 추천한다. 구매가능점수[3] pls(c,j)는 다음의 식 (5)와 같다.

$$pls(c,j) = \frac{\sum_{i \in H} p_{ij} \times cen(i)}{\sum_{i \in H} cen(i)} \quad (5)$$

p_{ij}는 고객 i의 j번째 상품에 대한 구매여부를 나타내며 cen(i)는 고객 i의 중심성값, H는 신규고객 이웃집합을 의미한다. pls값이 높으면 이웃 고객들의 구매빈도가 높다는 의미로 신규고객의 구매 확률도 높다는 것을 의미한다. 그러므로 pls값이 높은 상위 N개의 상품을 신규고객에게 추천한다. 기존 협력필터링은 고객의 구매기록이 있으므로 유사도 계산이 가능하지만, 신규고객의 경우 계산할 수 없으므로, 유사도가 아닌 중심성 값이 높은 기존고객들을 이웃고객으로 사용한다(박종학 외. 2009).

3) 구매가능성점수(pls) : purchase likelihood score

제 5 절 추천 및 비추천시스템

제 1 절에서 제 4 절까지 언급한 과정을 다음의 두 단계로 구분하여 분석하는데 이를 간략하게 표현하면 아래 < 표 3-2 >와 같다.

< 표 3-2 > 추천 및 비추천시스템

	Positive Centrality	Negative Centrality
Positive pls	중심성 : high pls : high	중심성 : low pls : high

중심성이 높은 고객들만으로 분석한 기존 연구들과는 다르게 본 논문에서는 중심성이 높은 고객의 영화목록 가운데 pls가 높은 영화, 중심성이 낮은 고객의 영화목록 가운데 pls가 높은 영화로 구분한다.

중심성이 높은 것을 Positive Centrality, 중심성이 낮은 것을 Negative Centrality라고 하고, pls가 높은 것을 Positive pls라고 한다. 즉, Positive Centrality와 Positive pls인 영화는 추천을 해야 하는 목록이다. 반대로 Negative Centrality와 Positive pls인 영화는 중심성이 낮은 고객들이 많이 선택한 영화이므로 비추천 목록이라고 볼 수 있다.

제 4 장 실증분석

제 1 절 분석 데이터

4.1 데이터

실험 분석을 위한 데이터로 MovieLens의 데이터를 사용한다. MovieLens는 1997년 9월부터 1998년 4월까지 웹 사이트를 통해 수집된 것으로 미네소타 대학의 GroupLens Research Project팀에 의해 수집된 영화추천시스템이다.

조사기간 동안 수집된 1682편의 영화에 대한 943명의 고객 평가 데이터이다. 고객ID, 영화ID, 각각의 고객들이 영화에 대해 1점에서 5점 사이의 점수를 부여한 평가치, 평가한 시간 등 총 100,000개의 데이터가 수집되었다.

전체 데이터를 시간 순으로 배열하여 초기 50,000개를 기존고객 데이터로, 후기 50,000개를 신규고객 데이터로 구분하였다. 강부식(2010)은 신규고객 데이터를 기존고객 데이터에 없는 고객으로 신제품에 해당하는 영화데이터는 제외하고 가장 먼저 본 영화 10편을 추출하였다. 신규고객 데이터는 기존고객 데이터를 활용한 분석의 검증을 위한 데이터로 활용하였다.

따라서 신규고객은 452명 각각의 영화데이터 10편으로 총 4,520개의 데이터로 구성하였다.

4.2 인구통계분석

전체 데이터에서의 응답자 연령은 다음 < 표 4-1 >과 같이 '21세~30세'가 35.95%(339)로 가장 높게 나타났으며 '71세~80세'가 0.11%(1)로 가장 낮게 나타났다.

< 표 4-1 > 응답자 연령

구분	빈도	누적빈도	유효비율	누적비율
10세 이하	2	2	0.21	0.21
11세~20세	107	109	11.35	11.56
21세~30세	339	448	35.95	47.51
31세~40세	223	671	23.65	71.16
41세~50세	167	838	17.71	88.87
51세~60세	83	921	8.80	97.67
61세~70세	21	942	2.23	99.89
71세~80세	1	943	0.11	100.00
합계	943	-	100.00	-

응답자의 성별은 다음의 < 표 4-2 >와 같이 '남자'가 71.05%(670), '여자'가 28.95%(273)의 비율인 것으로 나타났다.

< 표 4-2 > 응답자 성별

구분	빈도	누적빈도	유효비율	누적비율
여자	273	273	28.95	28.95
남자	670	943	71.05	100.00
합계	943	-	100.00	-

응답자의 직업은 다음의 표 < 표 4-3 >과 같이 나타났다. ‘student’가 20.78%(196)로 가장 높은 것으로 나타났으며, ‘doctor’가 0.74%(7)로 가장 낮은 빈도인 것으로 나타났다.

< 표 4-3 > 응답자 직업

구분	빈도	누적빈도	유효비율	누적비율
administrator	79	79	8.38	8.38
artist	28	107	2.97	11.35
doctor	7	114	0.74	12.09
educator	95	209	10.07	22.16
engineer	67	276	7.10	29.27
entertainment	18	294	1.91	31.18
executive	32	326	3.39	34.57
healthcare	16	342	1.70	36.27
homemaker	7	349	0.74	37.01
lawyer	12	361	1.27	38.28
librarian	51	412	5.41	43.69
marketing	26	438	2.76	46.45
none	9	447	0.95	47.40
other	105	552	11.13	58.54
programmer	66	618	7.00	65.54
retired	14	632	1.48	67.02
salesman	12	644	1.27	68.29
scientist	31	675	3.29	71.58

student	196	871	20.78	92.36
technician	27	898	2.86	95.23
writer	45	943	4.77	100.00
합계	943	-	100.00	-

아래 < 표 4-4 >는 응답자의 응답 지역을 빈도별로 나타낸 것이다.

< 표 4-4 > 응답자 지역

구분(주)	빈도	누적빈도	유효비율	누적비율
Alabama (AL)	3	3	0.32	0.32
Alaska (AK)	6	9	0.64	0.95
APO Europe	1	10	0.11	1.06
arizona	14	24	1.48	2.55
California (CA)	118	142	12.51	15.06
Colorado (CO)	20	162	2.12	17.18
Connecticut	17	179	1.80	18.98
Delaware	3	182	0.32	19.30
Florida (FL)	24	206	2.55	21.85
Georgia (GA)	21	227	2.23	24.07
Hawaii (HI)	4	231	0.42	24.50
Idaho (ID)	7	238	0.74	25.24
Illinois (IL)	17	255	1.80	27.04
Indiana (IN)	10	265	1.06	28.10
Iowa (IA)	16	281	1.70	29.80
Kansas 등	5	286	0.53	30.33
Kentucky (KY)	13	299	1.38	31.71
Louisiana (LA)	7	306	0.74	32.45
Maine	2	308	0.21	32.66
Maryland (MD)	44	352	4.67	37.33
Massachusetts	43	395	4.56	41.89
Michigan (MI)	24	419	2.55	44.43
Minnesota (MN)	81	500	8.59	53.02
Missouri (MO)	17	517	1.80	54.83

Montana (MT)	2	519	0.21	55.04
Nebraska (NE)	6	525	0.64	55.67
Nevada (NV)	3	528	0.32	55.99
New Hampshire	6	534	0.64	56.63
New Jersey	19	553	2.01	58.64
New Mexico (NM)	2	555	0.21	58.85
NEWPORTRICHEY	2	557	0.21	59.07
New York	61	618	6.47	65.54
North Carolina (NC)	19	637	2.01	67.55
North Dakota (ND)	2	639	0.21	67.76
Ohio (OH)	34	673	3.61	71.37
Oklahoma (OK)	9	682	0.95	72.32
Oregon (OR)	21	703	2.23	74.55
Pennsylvania	35	738	3.71	78.26
South Carolina (SC)	11	749	1.17	79.43
South Dakota (SD)	1	750	0.11	79.53
Standard	34	784	3.61	83.14
Tennessee (TN)	15	799	1.59	84.73
Texas (TX)	51	850	5.41	90.14
Utah (UT)	9	859	0.95	91.09
Vermont	6	865	0.64	91.73
Virginia (VA)	27	892	2.86	94.59
Washington (WA)	23	915	2.44	97.03
West Virginia (WV)	3	918	0.32	97.35
Wisconsin (WI)	24	942	2.55	99.89
Wyoming (WY)	1	943	0.11	100.00
합계	943	-	100.00	-

제 2 절 분석 방법

본 연구에서는 협력필터링과 사회연결망의 중심성 개념을 활용하여 신규고객에게 추천하는 방법과 비추천하는 방법을 각각 분석한다. 기존고객 데이터 50,000개에서 상관계수를 이용하여 유사도를 측정한다. 그 과정에서 0.2에서 0.8까지, -0.2에서 -0.8까지 각각 0.2씩 증가시켜 ρ 값을 변화시킨다. 이웃고객의 수는 50으로 하여 이웃고객들의 구매정보로 pls를 측정한다. pls 점수가 높은 순으로 10개의 영화를 추천 목록으로 생성하고, 이 영화를 신규고객에게 추천하여 추천정확성 평가를 평가한다.

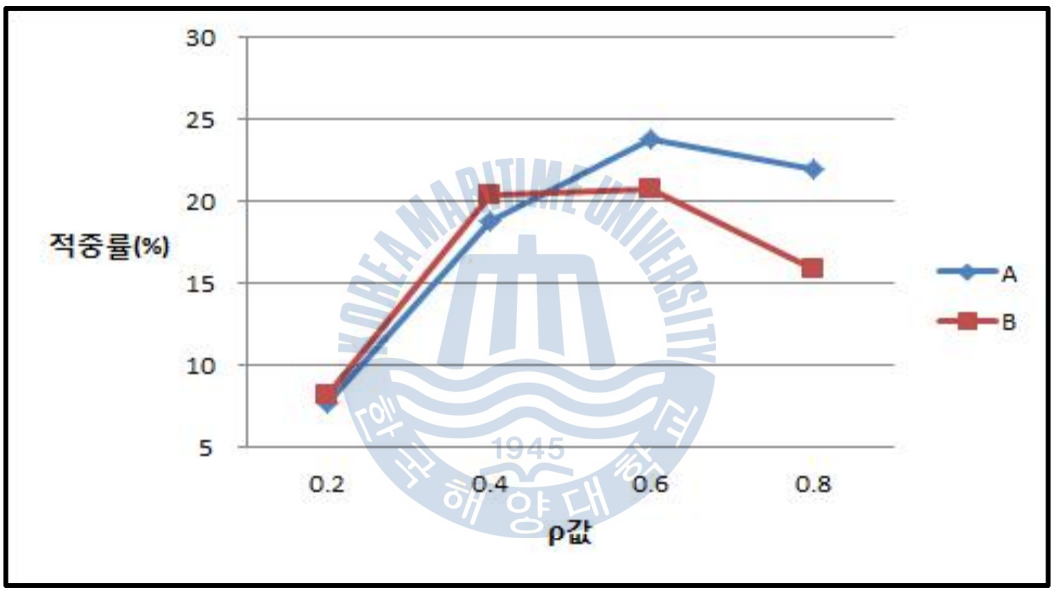
추천정확성 평가를 위해 적중률(신동원, 2006)을 이용한다.

$$\text{적중률}(\%) = \frac{\text{추천상품을구매한신규고객}}{\text{전체신규고객}} \times 100 \qquad (6)$$

분석을 위해 Visual Basic, UCINET, SPSS를 사용하였다. 사회연결망 분석을 위하여 본 논문에서는 UCINET을 사용하였는데 그 외에도 Pajek, NetMiner 등의 프로그램들이 사회연결망 분석을 위하여 사용되기도 한다.

제 3 절 분석 결과

아래의 < 그림 4-1 >은 Positive Centrality - Positive pls의 분석결과이다. 이웃 고객과 성향이 비슷하다는 측면에서 대표성을 띄는, 즉 중심성이 높은 집단에 포함되는 고객이 본 영화 가운데 빈도가 높은 것을 신규고객에게 추천하고 그 적중률을 나타낸 것이다. 기존고객 가운데 중심성이 높은 고객이 선호한 영화는 전체 고객과 영화 선호도에 대한 공통점이 많은 것으로 볼 수 있으므로 신규고객에게 추천하는 목록이 된다.

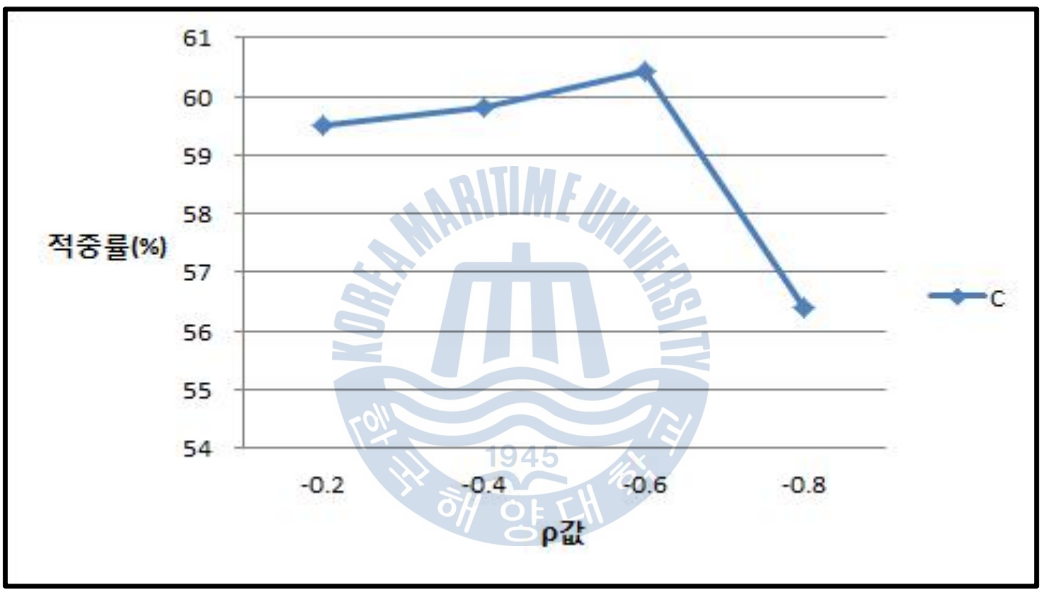


< 그림 4-1 > Positive Centrality - Positive pls

분석 결과 고객 간의 유사도는 ρ 값이 0.2에서 0.6까지는 증가하는 것으로 나타났으며 ρ 값이 최적인 값은 0.6인 것으로 나타났다. 유사도 값이 낮은 경우에는 이웃 간의 유사성은 떨어지나 네트워크 형성이 잘 된다. 반면, 유사도 값이 커질수록 이웃 간의 유사성은 높아지지만 개체수가 줄어들어 네트워크의 형성이 어렵다.

< 그림 4-1 >에서 A는 신규고객을 위한 추천시스템의 적중률로 기존고객의 데이터로 생성한 추천목록을 신규고객에게 추천한 결과이다. ρ 값을 0.2에서 0.8까지 증가시킬 때 변화하는 적중률이다.

B는 신규고객을 위한 추천 목록을 변화시킨 것으로 A에서 추천한 목록과 < 그림 4-2 >에서 추천한 목록 가운데 서로 중복되는 데이터는 제거한 후의 적중률이다. B는 ρ 값이 0.2와 0.4일 때 A보다 높은 것으로 나타났다. ρ 값이 낮은 경우 B의 적중률이 더 높게 나타나는 이유는 집단 내에 이질적인 특성을 가진 고객들이 포함되어 있기 때문이다.



< 그림 4-2 > Negative Centrality - Positive pls

< 그림 4-2 >는 Negative Centrality - Positive pls의 분석결과이다. C는 중심성이 낮은 고객이 본 빈도가 높은 영화의 목록을 신규고객에게 추천하고 그 적중률을 나타낸 것이다.

중심성이 낮은 고객은 이웃고객과 같은 영화에 대한 중복도가 낮은 고객으로 다시 말하자면 영화에 대한 성향이 타인과는 다른 사람이라고 볼 수 있다. 그러므로 본 논문에서는 여기서 생성된 추천 목록을 신규고객에 대한 비추천 목록으로 본다.

앞서 언급했던 바와 같이 유사도 값이 낮은 경우에 이웃 간의 유사성은 떨어지지만 네트워크 형성이 잘 되어 개체수가 늘어난다. 개체수가 많을수록 고객과 고객 간의 연결빈도가 높아지고, 다양한 개체가 연결 되었으므로 적중률도 낮아진다. 반면, 유사도 값이 커질수록 유사성은 더욱 강해지지만 연결되는 개체수가 줄어들어 네트워크의 형성이 어려워지는 현상이 발생한다. 따라서 C는 ρ 값이 -0.6일 때 적중률이 가장 높은 것으로 나타났다.

ρ 값이 -0.8일 때 적중률이 가장 낮은 것은 ρ 값이 낮아지면 이질적인 고객이 많이 포함되어 있고 이질적인 고객의 다양한 정보를 중심으로 생성한 추천목록은 신규고객에게 추천하였을 때 적중률이 낮아질 수밖에 없기 때문이다.

제 5 장 결 론

제 1 절 연구결과 요약 및 시사점

본 논문은 MovieLens의 데이터를 사용하여 전체 1,682편의 영화에 대한 고객 943명의 평가로 분석 하였다. 고객이 영화에 대해 1점에서 5점 사이의 선호도 점수를 부여한 평가치로 이루어져 있으며 총 100,000개의 데이터가 수집되었다. 100,000개의 데이터를 각각 50,000개씩 기존고객, 신규고객 두 그룹으로 나누어 분석하였다.

기존에 존재하는 신규고객 추천에 대한 문제점을 해결하기 위하여 협력필터링 추천방법과 사회연결망의 중심성 분석을 적용 하였다. 신규고객을 위한 추천 방법을 찾고, 그 과정에서 사용되지 않았던 중심성이 낮은 고객의 데이터를 통한 새로운 정보를 찾는 것에 그 목적이 있다.

기존 연구에서는 대부분 중심성이 높은 고객들의 데이터가 사용되었다. 그러나 본 논문에서는 중심성이 낮은 고객들의 정보를 이용하여 비추천시스템을 분석하였다.

베스트셀러기반 추천방법은 모든 고객에게 같은 정보를 제공하므로 개인화된 추천이 불가능하며, 판매가 많이 된 상품이 특정 상품 군에 속하는 경우가 많으므로 추천 정확도가 낮다. 그러나 본 논문에서 제시한 추천방법은 유사도가 높은 기존고객 간의 관계망 형성을 이용하여 신규고객을 위한 추천이 가능하다는 장점이 있다.

또한 신규고객 뿐만 아니라 기존고객에게 추천을 할 때도 고객의 과거 구매 데이터를 기초로 하여 해당 기존고객이 Positive Centrality - Positive pls와 Negative Centrality - Positive pls중 어디에 속하는지를 파악한 후, 성향에 맞는 적절한 추천의 제공이 가능해진다.

따라서 고객의 니즈에 부합하는 제품을 추천하여 만족도를 증가시켜 고객의 충성도를 보다 높이고 CRM의 측면에 기여할 것이다.

제 2 절 연구의 한계 및 향후 연구방향

본 논문은 GroupLens Research Project팀에 의해 구성된 영화추천시스템에서 수집된 정보로 다른 상품에 대한 추천시스템으로 사용하는 것에는 상품의 특이성으로 인한 한계가 존재할 수 있다. 고객의 과거 구매 데이터를 이용하여 추천목록을 작성하여 임의로 가정한 신규고객에게 추천하는 방식을 사용하였으므로 다양한 관심사가 존재하는 실제 고객들에게 모두 적용하는 것도 한계가 존재한다.

시간 순으로 나누어 분석하였지만 기존고객 그룹과 신규고객 그룹의 특성이 유사하다고만 볼 수는 없으므로 기존고객 데이터로 신규고객에게 추천하는 것이 완벽하게 적합하다고 볼 수 없는 부분도 한계점이다.

따라서 추천시스템을 적용 할 때에는 각 상품의 특성이나 상황에 적합하게 이웃 고객 수, 추천 목록 수, ρ값 등을 설정할 필요가 있다.

또한 사회네트워크의 중심성만을 사용하여 분석하였기 때문에 구조적 공백, 밀도 등 전반적인 부분을 고려한 분석을 한다면 정확도가 더욱 높아 질 것으로 예상된다.

참고문헌

◦ 강부식(2010), “사회연결망의 구조적 공백을 활용한 신규고객 웹 상품추천방법,” 산업경제연구, 제23권 제5호, pp.2371-2385.

◦ 강부식(2010), “사회연결망을 활용한 신규고객 상품추천방법의 추천정확성 향상,” Proceedings of the Korean Data Analysis society, pp.53-61.

◦ 국승용(2007), “연결망 분석기법을 활용한 농산물 물류센터의 입지특성 분석,” 농촌경제, 제30권 제4호, pp.221-235.

◦ 김용학(2003), 「사회 연결망 분석」, 박영사.

◦ 김용학(2003), 「사회 연결망 이론」, 박영사.

◦ 김재경, 조윤호, 김승태, 김혜경(2005), “모바일 전자상거래 환경에 적합한 개인화된 추천시스템,” 경영정보학연구, 제15권 제3호, pp.223-241.

◦ 박종학, 조윤호, 김재경(2009), “사회연결망 : 신규고객 추천문제의 새로운 접근법,” 지능정보연구, 제15권 제1호, pp.123-140.

◦ 배순한, 서재교, 백승익(2010), “온라인 커뮤니티의 중심성 변화에 대한 탐색적 연구:사회연결망 분석을 이용하여,” 지식경영연구, 제11권 제2호, pp.17-35.

◦ 손동원(2002), 「사회 네트워크 분석」, 경문사.

◦ 신동원(2006), “사회연결망 분석을 활용한 신규고객 추천 방법론,” 국민대학교 석사학위논문.

◦ 이기헌, 김창욱(2010), “신제품 확산 다이나믹스 예측: 심리적 편향과 동적 구매

한계점을 고려한 다중 행위자 시뮬레이션," 대한산업공학회 추계학술대회논문집.

◦ 이용준, 이세훈, 왕창종(2003), "인구 통계 정보를 이용한 협업 여과 추천의 유사도 개선 기법," 정보과학회논문지 : 컴퓨팅의 실제, 제9권 제5호, pp.521-529.

◦ 이재식, 박석두(2007), "장르별 협업필터링을 이용한 영화 추천 시스템의 성능 향상," 한국지능정보시스템학회논문지, 제13권 제4호, pp.65-78.

◦ 이지선, 강신겸(2010), "커뮤니티관광개발에서의 이해관계자간 협력관계 분석 -사회연결망 분석을 중심으로-," 관광연구농촌, 제22권 제2호, pp.75-97.

◦ Bonacici, P.(1987), “Power and Centrality : A Family of Measures,” American Journal of Sociology, Vol. 92, pp.1170-1182.

◦ Burt, R.S.(1892), 「Structural Holes: The Social Structure of Competition」, Harvard University Press.

◦ Kohrs, A. and Merialdo, B.(2001), “Creating User-adapted Websites by the Use of Collaborative Filtering,” Interacting Computers, Vol. 13, pp.695-716.

◦ Lee, C. H., Kim, Y. H. and Rhee, P. K.(2001), “Web personalization Expert with Combining Collaborative Filtering and Association Rule Mining Technique,” Expert Systems with Applications, Vol. 21, pp.131-137.

◦ Sarwar, B., Karypis, G., Konstan, J. and Riedl, J.(2000), “Analysis of Recommendation Algorithms for E-Commerce,” Proceedings of ACM E-commerce 2000 conference, pp.158-167.

◦ Schein, A. I., Popescul, A., Ungar, L. H. and Pennock D. M.(2002), “Methods and Metrics for Cold-Start Recommedations,” SIGIR’02.

◦ Yu, K., Schwaighofer, A., Tresp, V. and Kriegel, H.(2004), “Probabilistic Memory-based Collaborative Filtering,” IEEE Transactions on Knowledge and Data Engineering, Vol. 16, pp.56-69.