ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2017 December: 23(4): 111~126 http://dx.doi.org/10.13088/jiis.2017.23.4.111

사회연결망 분석을 활용한 연관규칙 확장기법

이동원
한성대학교 경영학부
(*dongwonlee@hansung.ac.kr*)

연관 상품 추천은 수많은 상품을 다루는 온라인 상거래에서 소비자의 상품 탐색 시간을 줄여주며 판매자의 매출 증대에 크게 기여한다. 이는 주문과 같은 거래의 빈도를 기반으로 생성되므로, 통계적으로 판매 확률이 높은 상품을 효과적으로 선별할 수 있다. 하지만, 판매 가능성이 높은 경우라도 신상품처럼 판매 초기에 거래 건수가 충분하지 않은 상품은 추천에서 누락될 수 있다. 연관 추천에서 누락된 상품은 이로 인해 노출 기회를 잃게 되고, 이는 거래 건수 감소로 이어져, 또 다시 추천 기회를 잃는 악순환을 겪을 수도 한다. 따라서, 충분한 거래 건수가 쌓이기 전까지 초기 매출은 일정 기간 동안 정체되는 현상을 보이는데, 의류 등과 같이 유행에 민감하거나 계절 변화에 영향을 많이 받는 상품은 이로 인해 매출에 큰 타격을 입을 수도 있다.

본 연구는 이와 같이 거래 초기의 낮은 거래 빈도로 인해 잘 드러나지 않는 상품 간의 잠재적인 연관성을 찾아 추천 기회를 확보할 수 있도록 연관 규칙을 확장하기 위한 목적으로 수행되었다. 두 상품 간에 직접적인 연관성이 나타나지 않더라도 다른 상품을 매개로 두 상품 간의 잠재적 연관성을 예측할 수 있을 것이며, 이런 연관성은 주문에서 나타나는 상품 간 상호작용으로 표현될 수 있으므로, 사회연결망 분석을 활용한 분석을 시도하였다. 사회연결망 분석기법을 통해 각 상품의 속성과 두 상품 간 경로의 특성을 추출하고 회귀분석을 실시하여, 두 상품 간 경로의 최단 거리 및 경로의 개수, 각 상품이 얼마나 많은 상품과 연관성을 갖는지, 두 상품의 분류 카테고리가 어느 정도 일치하는지가 두 상품 간의 잠재적 연관성에 미친다는 것을 확인하였다.

모형의 성능을 평가하기 위해, 일정 기간의 주문 데이터로부터 연결망을 구성하고, 이후 10일 간 생성될 상품 간 연관성을 예측하는 실험을 진행하였다. 실험 결과는 모형을 적용하지 않는 경우보다 제안 모형을 활용할 때 훨씬 많은 연관성을 찾을 수 있음을 보여준다.

주제어 : 추천 시스템, 연관 규칙 마이닝, 사회 연결망 분석, 연관 규칙 확장, Cold Start Problem

논문접수일 : 2017년 7월 31일 논문수정일 : 2017년 9월 17일 게재확정일 : 2017년 9월 20일
원고유형 : 일반논문 교신저자 : 이동원

1. 서론

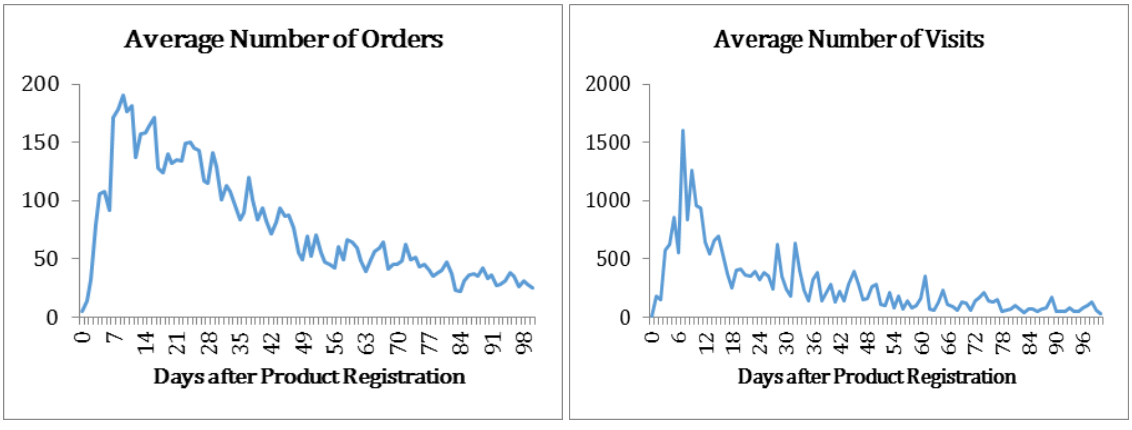
수많은 상품을 다루는 온라인 상거래에서 소비자가 원하는 상품을 빠르게 찾고 이를 구매하도록 돕기 위해 추천 시스템이 활용되고 있다. 웹 페이지라는 가상의 진열 공간의 활용은 물리적 공간의 제약을 극복할 수 있게 해줌으로써 오프라인 매장에서와는 비교할 수 없는 수많은 상품들을 진열할 수 있게 되었다. 이로 인해, 소비자는 많이 팔리는 대중적인 상품이 아니더라도 자신만의 취향에 더 잘 맞는 상품을 구매할 수 있게 되었다. 하지만, 온라인에 진열된 수많은 상품 중 자신이 원하는 상품을 찾기 위해 비교하는 노력은 크게 증가하게 되었다. 소비자의 상품 탐색을 돕기 위해 온라인 기업은 상품 추천 기능을 제공하고 있는데, 이는 대안의 폭을 줄여줌으

로써 소비자가 좀 더 쉽게 상품을 선택하도록 돕는다.

연관 상품 추천 기능은 소비자가 관심을 보인 특정한 상품에 밀접하게 관련된 상품들을 보여줌으로써, 다른 대안으로 비교 가능한 유사 상품 또는 추가적으로 구매 가능한 상품을 제공하는 수단으로 활용되고 있다. 하지만, 이런 연관 상품 추천 기법은 동시 혹은 순차적인 주문과 같이 이미 발생한 거래를 통해 상품 간의 직접적인 연관성이 드러난 상품만을 대상으로 한다는 점에서, 연관성이 높지만 주문 빈도가 낮은 상품 간의 연관성을 찾기 힘들다는 문제를 갖는다. 특히, 신규로 등록된 상품의 경우에는 누적된 거래 건수가 작아 다른 상품에 비해 상대적으로 노출 기회를 확보하기 어렵다. 추천 노출 빈도의 감소는 다시 거래의 감소를 야기시키는 상황을 야기시키는데, 이는 어느 정도 거래의 수가 만들어지기 전까지는 해소되지 않는 소위 콜드 스타트(Cold Start) 문제를 발생시킨다. <Figure 1>은 신규상품이 등록된 날짜로부터 주문 건수와 상품 페이지 방문 건수의 평균값이 각각 어떻게 변화하는지를 보여준다. 이로부터 상품이 등록된 날로부터 평균적으로 10일 정도가 지난 후에야 주문과 상품 페이지 방문이 크게 높아지며 거래가 활성화되는 것을 확인할 수 있다.

초기 노출 기회의 상실은 상품의 판매 주기에 걸친 매출의 감소로 이어질 수 있다. 특히, 유행에 민감한 상품이거나 계절의 변화가 매출에 민감하게 영향을 미치는 상품의 경우, 초기 노출이 원활하지 않을 경우 실질적인 판매 기간이 감소되는 효과로 인해 매출에 큰 타격을 입을 수 있다. 신상품의 노출 기회를 상대적으로 늘려주는 방안도 고려해볼 수 있으나, 실제로 연관성이 낮은 상품도 함께 노출됨으로 인해 연관 추천의 신뢰성을 떨어뜨리는 역효과를 일으킬 수 있다. 따라서, 거래 빈도가 낮은 상품 중 잠재적으로 연관성이 높을 상품을 선별할 수 있는 기법에 관한 연구가 필요하다.

이에 본 연구는 거래량이 부족한 경우에도 연관 상품 추천이 이루어질 수 있도록 사회연결망을 활용하는 방안을 제안한다. 사회연결망 분석은 노드로 표현되는 개체와 개체 간의 상호작용



(a) Orders (b) Visits

〈Figure 1〉 Changes in the Average Number of Transactions

을 링크로 표현하는데, 연관 상품의 경우에도 한 상품의 주문이 다른 상품의 주문에 영향을 미치는 것을 상품 간 상호작용으로 이해할 수 있다. 따라서, 본 연구는 주문을 기반으로 사회연결망을 구성하고, 연결망 내에서 서로 연결된 상품 간 연결 속성을 분석하며, 이를 기반으로 아직 드러나지 않은 상품 간 연관성을 발견하는 방법을 체계적으로 제시하여, 기존의 연관 규칙 마이닝 기법의 한계를 극복하는 방안을 제시한다. 제안된 모형의 성능은 특정 시점을 기준으로, 이 시점 이전의 연결망에서 상품 간의 연결 특성으로부터 이후에 나타날 잠재적 연관성을 예측하는 방법으로 평가하였다. 실험 결과로, 모형에 의해 잠재적 연관성을 갖는 것으로 예측된 상위 10%에서 평균 예측치의 4배 이상 우수한 성능을 보이는 것으로 나타났다.

이후의 구성은 다음과 같다. 2장에서는 이론적 배경을 설명하고, 3장에서는 연구 모형의 설계 과정을 살펴본다. 제안된 모형의 성능을 평가한 결과를 4장에서 정리하고, 5장에서 결론으로 마무리한다.

2. 이론적 배경

2.1 추천 시스템

온라인 상거래에서 추천 시스템은 상품에 대해 사용자가 표현한 선호도를 기반으로 서로 다른 상품 간 유사도나 서로 다른 사용자 간 유사도를 계산하고 이를 기반으로 사용자가 아직 경험하지 못한 품목(상품 또는 서비스)을 추천하는 일종의 의사결정지원 시스템이다. 사용자의 관점에서 추천 시스템은 자신이 원하는 품목을 찾기 위한 탐색 노력을 줄여주는 역할을 수행하고, 기업에게는 고객 충성도와 함께 매출을 증대하는 효과를 가져온다(Ansari et al., 2000). 이와 같은 추천 시스템에 대해서는 수많은 연구가 수행되고 있다. 더 많은 추천이 사용자로부터 수용되도록 하기 위해 추천 기법의 성능을 높이기 위한 연구(Balabanovic and Shoham, 1997; Ansari et al., 2000; Adomavicius and Tuzhilin, 2011; Choi et al., 2016)가 지속적으로 이뤄지고 있다. 또 한편으로는, 추천 시스템의 성과를 측정하기 위한 연구(Bodapati, 2008; Fleder and Hosanagar, 2009), 그리고 상거래 이외의 다양한 분야에서 활용하기 위한 연구(Choi et al., 2015; Kim and Lee, 2013; Kim et al., 2010)가 활발히 수행되고 있다.

추천 시스템의 성능을 높이기 위한 기법으로, 내용기반 필터링(content-based filtering)과 협업 필터링(collaborative filtering)이 주목 받고 있다. 내용기반 필터링 기법은 서로 다른 품목 간의 유사성을 기반으로 적합한 추천 품목을 찾기 위해 의사결정나무, 최근접 이웃 기법 등 다양한 분류 기법을 활용한다(Konstan et al., 1997; Ansari et al., 2000). 반면, 협업 필터링은 서로 다른 사용자 간의 유사성을 기반으로, 유사 사용자가 선호한 품목 중 추천 대상 사용자가 아직 경험하지 않은 상품을 추천하는 방식이다(Konstan et al., 1997; Ansari et al., 2000).

2.2 연관 규칙 마이닝

Agrawal et al. (1993)이 제시한 연관 규칙 마이닝 기법은 주문과 같이 반복적으로 발생하는 거래에서 반복적으로 함께 출현하는 품목 간의 연관성을 패턴으로 표현하는 방법을 일컫는다. 이

런 패턴은 관련된 서로 다른 품목 간의 연관성의 강도를 포함하는 연관 규칙의 형태를 갖는데, 연관성의 척도로는 두 품목이 동시에 같은 거래에 등장하는 빈도가 사용된다. 주문에서 나타난 품목 간의 연관성은 어떤 품목의 판매가 또 다른 품목의 판매 가능성을 암시한다고 할 수 있다. 따라서, 이를 정형화한 연관규칙은 온라인 상거래에서 개별 상품 페이지에 추가로 판매가 될 가능성이 높은 상품을 추천하기 위해 활용되고 있다(Anand, 1998; Chen et al., 2006; Kim and Street, 2004; Lee et al., 2013; Kim and Kim, 2005).

연관 규칙은 아래와 같이 정형화된 형태를 갖는데, A와 C는 각각 선행 품목(antecedent item)과 후행 품목(consequent item)을 sup와 conf는 지지도(support)와 신뢰도(confidence)를 의미한다.

$$A \rightarrow C\ (\sup\%, \mathrm{conf}\%)$$

지지도와 신뢰도는 연관 규칙에 포함된 선후행 품목 간 연관성의 강도를 나타내는 척도로서 사용된다. 여기서, 지지도는 전체 거래의 건수 중 선행 품목과 후행 품목이 동시에 나타난 거래 건수의 비율로 측정되며, 신뢰도는 선행 품목이 나타난 거래 건수 중 후행 품목이 함께 포함된 거래 건수의 비율, 즉 조건부 확률로 계산된다. 선행 품목을 구매했거나 이에 관심을 보인 소비자는 이와 연관된 여러 후행 품목 중 신뢰도가 높은 품목일수록 더 흥미를 보일 가능성이 높을 것으로 기대될 수 있다. 현업에서는 이런 상품으로부터 추천 목록을 작성하고 이를 선행 품목의 소개 페이지에 노출하는 방법으로 매출을 높이기 위해 노력하고 있다.

선행 상품과 후행 상품이 함께 포함된 거래가 사전에 충분히 발생하지 않게 되면 두 품목 간의 낮은 지지도로 인해 후행 상품은 선행 상품의 추천 목록에 포함되지 못한다. 하지만, 이런 상황은 후행 품목이 선행 품목과 함께 주문될 기회를 감소시켜 두 품목을 포함하는 연관 규칙의 지지도를 더 낮추는 결과로 이어진다. 더욱이, 새로 등록된 품목의 경우 그 자신의 주문 건수가 낮아 연관 상품으로 추천되기 더욱 힘들다는 문제(Cold Start Problem)가 제기된다.

2.3 사회연결망(Social Network)

사회 연결망은 개인이나 조직과 같은 사회적 개체와 이들 간의 상호작용으로 구성된 사회적 구조를 일컫는다. 이는 개체의 개별 속성이 아닌 개체 간의 관계를 이해하려는 시도를 일컫는다(Yun and Chae, 2005; Sohn, 2002; Kim, 2003). 이에 대한 연구는 개체를 노드(Node), 이들 간의 관계를 링크(Link)로 표현하는 연결망(Network)에서 이들 간의 연결 상태 및 연결 구조를 계량적으로 측정하고 시각적으로 표현하는 사회연결망 분석(Social Network Analysis) 기법을 활용한다. 사회연결망 분석에서는 연결망의 특성을 파악하기 위해, 밀도(Density), 중심성(Centrality), 중심화(Centralization)과 같은 척도를 활용한다. 밀도는 연결망 내에서 노드 간에 얼마나 많은 링크가 연결되었는가를 판단하는 척도로서, 연결 가능한 링크의 수에 대해 실제로 연결된 링크의 수를 계산한다. 중심성은 각 노드가 연결망 내에서 얼마나 중심적인 역할을 수행하는가를 판단하는 척도로서, 대표적으로 연결 정도 중심성(Degree Centrality), 근접 중심성(Closeness Centrality), 매개 중심성(Betweenness Centrality) 등이 활용되고 있다. 한 노드가 다른 노드와 연결된 정도를 판단하는 척도로서, 그 연결의 수가 많을수록 높은 값을 갖는다. 근접 중심성은 한

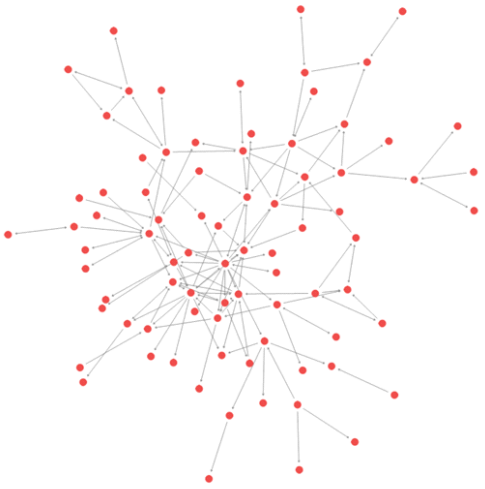
노드가 다른 노드와 얼마나 가깝게 연결되어 있는가를 판단하는 척도로서, 다른 노드에 이르는 거리가 짧을수록 높은 값을 갖는다. 매개 중심성은 한 노드가 다른 노드들 간을 연결하는 역할을 수행하는 수준을 판단하는 척도로서, 노드 간의 최단 경로에 위치하는 비율이 높을수록 높은 값을 갖는다. 중심화는 연결망이 특정 노드를 중심으로 얼마나 집중되어 있는지를 판단하는 척도로서, 연결 정도 집중도(Degree Centralization), 근접 집중도(Closeness Centralization), 매개 집중도(Betweenness Centralization)가 주로 활용되고 있다. 이들 각각은 연결 정도, 근접도, 매개 수준을 근거로 계산된다.

사회연결망을 추천 시스템에 적용하고 시도한 연구(Kim and Chang, 2010; Kim and Kim, 2014; Kim and Kim, 2016; Noh et al., 2017; Kang, 2010; Kim et al., 2010; Shin et al., 2012; Part et al., 2009)는 다양하게 수행되었으나, 기존 연구들은 주로 개인화 추천에 초점을 맞추고 있으며 연관 규칙을 확장하고자 한 연구는 찾아보기 힘들다. 그러나, 현업에서는 개인의 특성을 파악한 맞춤형 추천과 더불어 특정 상품에 관심을 보인 불특정 다수를 대상으로 한 연관 추천이 매출에 크게 기여하고 있다는 점에서, 연관 추천을 위한 사회연결망 활용방안에 과한 연구가 진행되어야 할 필요가 있다.

3. 연구 모형

3.1 데이터 수집

본 연구에서는 온라인 상거래 기업으로부터 수집한 실제 주문 거래 데이터를 사용한다. 데이터의 수집 기간은 2016년 2월부터 4월까지 3개월이며, 분석의 용이성을 높이기 위해 잡화 카테고리에 한정하여 분석을 실시하였다. 분석에 사용된 잡화 카테고리의 상품의 수는 932개이며, 이를 구매한 고객의 수는 모두 18,410명이었다. 이는 네트워크의 규모가 너무 크면 분석의 복잡성이 높아지기 때문이다. 사회연결망 구성에 필요한 데이터를 확보하기 위하여, 각 고객별로 주문한 상품을 시간 순으로 나열하여 선행상품(A)과 후행상품(C)을 포함하는 659개의 고유한 순서쌍(pair)을 추출하였다. 고객 중 한 개의 상품만을 구매한 경우에는 순서쌍을 작성할 수 없어 고객 중 16,825명이 제외되고 2건 이상의 거래가 존재하는 1,585명에 대해서만 순서쌍을 추출하여, 이들 사이의 링크(A → C)를 기반으로 <Figure 2>와 같이 상품 주문 연관성 네트워크를 구성하였다. 두 상품의 주문은 순차적으로 발생하므로 주문 순서에 따라 링크는 방향성을 갖도록(Directed) 설정하였으며, 가중치(Weight)는 1을 부여하였다. 연결망의 분석 도구로는 넷마이너4를 활용하였다.



〈Figure 2〉 Social Network of Association Rules

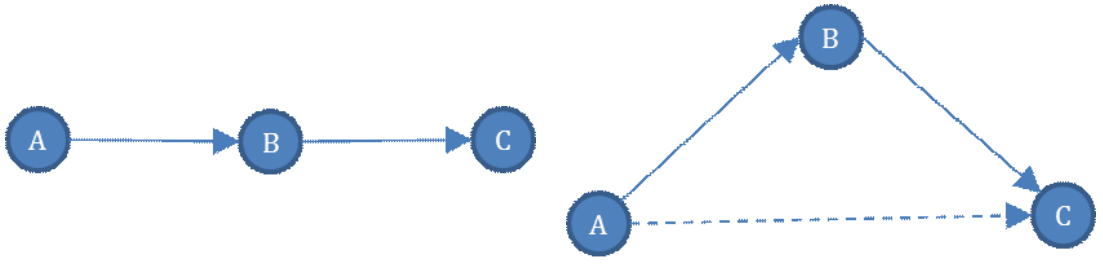
3.2 모형 설계 및 분석

거래 빈도가 낮아 아직 연관성이 밝혀지지 않은 두 개의 품목 간에 연관성을 찾기 위해 본 연구는 사회 연결망 분석기법을 적용한 모형을 고안한다.

예를 들어, 세 개의 품목 A, B, C에 대해 장바구니를 분석한 결과, A와 B는 동일 장바구니에서 발견되어 연관 규칙 A → B가 생성되었고, 마찬가지로, B와 C에 대해서도 연관 규칙 B → C가 발견되었으나, A와 C는 동일 장바구니에서 발견되지 않아 이들 간의 연관 규칙은 생성되지 않은 상황을 가정한다. 이를 연결망으로 표현하면 <Figure 3>의 (a)와 같다.

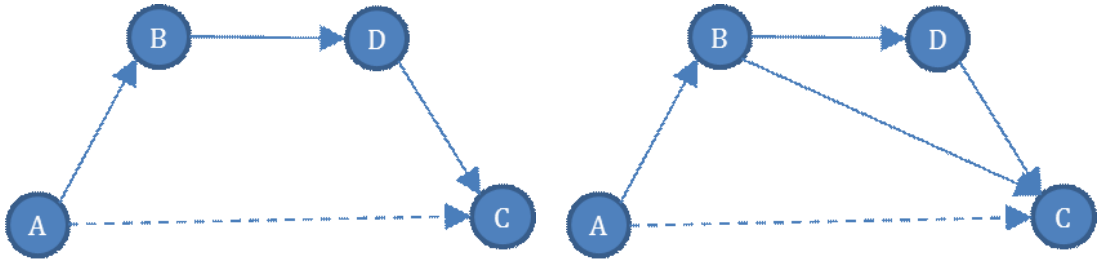
만약, 연관 규칙을 분석한 결과 A와 B의 연관성이 매우 강하고, B와 C 또한 강한 연관성을 띄게 된다면, A와 C 간에도 잠재적인 연관성이 있다고 기대할 수 있을 것이며, 이는 <Figure 3> (b)와 같이 표현 가능하다.

다음으로, 이렇게 유도된 잠재적 연관성의 강도에 영향을 미치는 요소에 대해 고려해보도록 하겠다. 우선 A와 C 사이에 다른 여러 품목을 거쳐야 하는 경우 둘 사이의 연관성은 낮을 것으로 기대할 수 있을 것이다. <Figure 4> (a)에서는 A에서 C에 이르기까지 두 개의 링크 A → B, B → D를 거쳐야 하므로 <Figure 3> (b)보다 A, C 간의 연결 강도가 약할 것으로 기대할 수 있을 것이다. 그러나, <Figure 4> (b)처럼, A → B → C와 A → B → D → C라는 두 개의 경로가 존재하는 경우, <Figure 3> (b), <Figure 4> (a) 어



(a) Association Rule Network (b) Derived Association Rule Network

〈Figure 3〉 Social Network of Two Association Rules



(a) Association Rule Network with One Path (b) Association Rule Network with Two Paths

〈Figure 4〉 Comparison of Association Rule Networks

느 쪽보다도 더 많은 경로를 가지므로 A → C 간 잠재적 연관성은 더욱 강할 것으로 기대할 수 있을 것이다.

이처럼, 두 개의 품목 간 잠재적 연결 강도는 둘을 간접적으로 이어주는 연결의 수와 둘 사이의 가장 빠른 경로를 통해 예측 가능할 것으로 기대할 수 있으므로, 최단 경로(Shortest Path)와 노드 연결성(Node Connectivity)를 사용하고자 한다. 각각은 사회연결망 분석에 사용되는 척도로서, 최단 경로는 두 노드 간의 가장 짧은 경로에 놓인 링크의 수를 의미하며, 노드 연결성은 두 노드 간의 연결이 완전히 끊어지게 하기 위해 제거해야 하는 링크 수로 정의된다. 이 값은 경로의 수가 많을수록 커지는 특성을 지니므로 본 연구에 적합하다고 판단된다.

또한, A, C가 각각 다른 노드들과 연결을 얼마나 잘 맺는지의 특성이 둘 사이의 연결 강도에 영향을 미친다고 기대할 수 있으므로, 각 노드의 중심성(Centrality)을 고려하기로 한다. 링크가 시작되는 A의 경우 C까지 연결되기 위해서는 가능한 한 많은 진출 차수(Out-Degree)를 갖는 것이 유리할 것이며, 반대로 C의 경우에는 진입 차수(In-Degree)의 영향을 받을 것으로 기대되므로 진출 차수 중심성(Out-Degree Centrality)과 진입 차수 중심성(In-Degree Centrality)를 예측 변수에 포함하기로 한다. 또한, 두 상품이 속한 카테고리는 두 상품 간의 유사성을 보여준다는 점에서 예측 변수에 포함하였다. 온라인 상품의 카테고리는 대분류, 중분류, 소분류, 세분류의 4단계를 갖는 계층적 구조를 갖는데, 낮은 카테고리에 속할수록 상품 간의 유사성이 높다. 이를 고려한 모형은 다음과 같다.

여기서, LinkWeight는 신규로 생성된 두 상품 간 링크의 수, Dist는 두 상품 간 최단 거리(Shortest Path), Conn는 두 상품 간 연결성(Node Connectivity), OutDegCent는 선행 상품의 진출 차수 중심성 (Out-Degree Centrality), InDegCent는 후행 상품의 진입 차수 중심성 (In-Degree Centrality), GroupMatch는 두 상품 간 카테고리 일치 수준(1: 대분류 일치, 2: 중분류 일치, 3: 소분류 일치, 4: 세분류 일치)을 의미한다. 본 연구에 사용된 변수는 Netminer4를 통해 측정되었으며, <Table 1>과 <Table 2>는 각각 이들 변수의 기술통계량과 변수 간 상관계수를 보여준다. <Table 2>에서 볼 수 있듯이 독립변수들 간의 상관관계가 낮아 이들을 모형에 사용하는 데에는 문제가 없다고 판단된다.

<Table 3>은 상품과 이들 간의 주문 연관성으로 구성된 네트워크에서 종속변수인 두 상품 간 연관성의 강도를 예측하는 모형의 회귀분석 결과이다. 모형에 사용된 설명변수는 모두 유의한 것으로 확인되었다. 즉, 두 상품 노드 간 최단 거리가 짧을수록, 두 노드 간 연결성이 높을수록, 또한, 선행 상품의 진출 차수 중심성이 높을수록, 후행 상품 진입 중심성이 높을수록, 두 상품의 분류 카테고리가 서로 가까울수록 더 많은 링크가 맺어질 가능성이 높다는 것으로 나타났다.

$$LinkWeight = \beta_0 + \beta_1 Dist + \beta_2 Conn + \beta_3 OutDegCent + \beta_4 InDegCent + \beta_5 GroupMatch + \varepsilon \quad (1)$$

〈Table 1〉 Descriptive Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
LinkWeight	9840	0.0381633	0.2493492	0	4.219508
Dist	9840	8.542683	3.153009	1	11
Conn	9840	1.175813	0.8196343	0	9
OutDegCent	9840	0.0238897	0.03409	0	0.232332
InDegCent	9840	0.0242664	0.0212238	0	0.0903841
GroupMatch	9840	1.639634	1.016589	1	4

〈Table 2〉 Correlation Coefficients

	LinkWeight	Dist	Conn	OutDegCent	InDegCent	GroupMatch
LinkWeight	1					
Dist	-0.348	1				
Conn	0.2437	-0.3475	1			
OutDegCent	0.206	-0.3998	0.2846	1		
InDegCent	0.123	-0.1881	0.3344	-0.0044	1	
GroupMatch	0.2205	-0.1689	0.1609	0.0659	0.0554	1

〈Table 3〉 Regression Analysis of Experiment Data

LinkWeight	Coef.	Std. Err.	t	P>t	[95% Conf. Interval]	
Dist	-0.020112	0.0008378	-24.01	0.000	-0.0217543	-0.0184697
NodeConn	0.0305911	0.0032138	9.52	0.000	0.0242915	0.0368908
OutDegCent	0.4798415	0.0754506	6.36	0.000	0.3319427	0.6277402
InDegCent	0.3907402	0.1165285	3.35	0.001	0.1623203	0.61916
GroupMatch	0.0380566	0.0023082	16.49	0.000	0.033532	0.0425812
Constant	0.0906602	0.0108173	8.38	0.000	0.0694562	0.1118643

4. 성능 평가

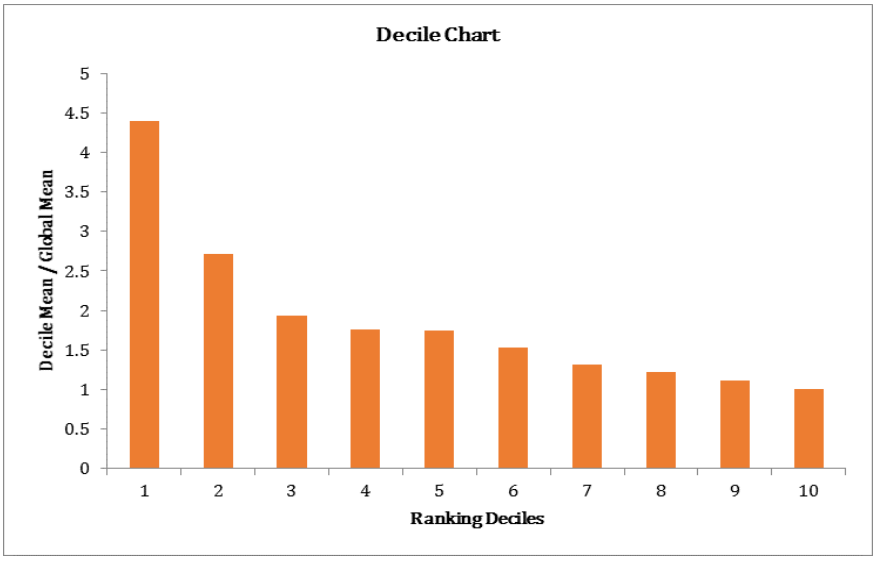
3장에서 분석한 모형에 의한 성능은 다음과 같이 평가한다. 연결망의 노드 중 서로 직접적으로 연결되지 않은 링크 중 이후에 연결될 링크의 수를 예측하는 것으로 평가한다. 이를 위해 데이터 수집 기간 중 마지막 10일 간 주문 데이터를 제외하고 연결망을 구성한 후, 제안 모형이 연결되어야 할 링크를 얼마나 찾아내는지를 확인하는 것으로 평가를 실시한다. 마지막 10일 이전(t-10일)에 작성된 연결망에서 직접 연결되지 않은 링크의 수는 5,711개였으며, 이들 중 이후 10일 간 새로 연결된 링크의 수는 611개였다. t-10일 시점의 네트워크에서 연결망 분석을 통해 5,711개의 각 잠재 링크 별로 설명변수를 추출하고, 모형을 통해 이 값으로부터 이후 10일 간 실제로 연결될 링크 수를 예측한다. 이렇게 계산된 값 중 가장 큰 값으로부터 가장 작은 값에 이르

〈Table 4〉 Result of Performance Experiment

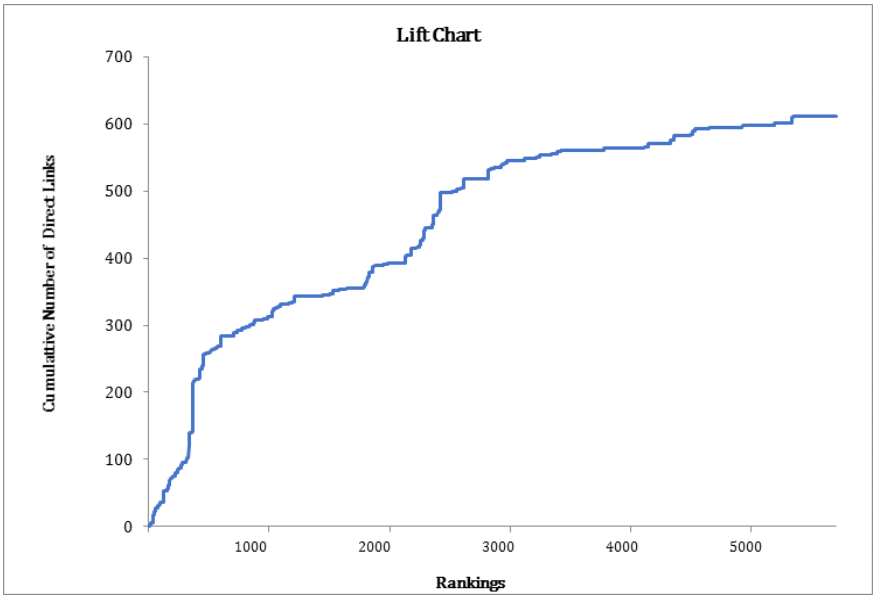
Decile	Number of Links	Cumulative Number of Links	Decile Mean / Global Mean
1	269	269	4.402618658
2	62	331	2.708674304
3	24	355	1.936715767
4	74	429	1.755319149
5	106	535	1.751227496
6	26	561	1.530278232
7	2	563	1.316343231
8	32	595	1.217266776
9	16	611	1.111111111
10	0	611	1

는 순서대로 실제로 10일간 연결된 611개의 링크와의 일치 여부를 확인하였다. 예측된 링크의 개수에 대해 실제 링크의 개수를 10분위 단위로 <Table 4>에 표기하였다.

모형에 의해 링크의 개수가 가장 많을 것으로 예측된 상위 10%인 571개(=5711*10%)의 잠재 링크 중 269개가 실제 링크로 확인되었으며 이는 모형 없이 평균적으로 발견 가능한 개수인 61개(=611*10%)에 비해 4.4배인 269개를 정확하게 예측할 수 있고, 20%인 1,142개를 예측할 경우에는 평균 개수 122개의 2.7배인 331개를 예측할 수 있는 것으로 확인되었다. 이를 <Figure 5>의 십분위 향상 차트와 <Figure 6>의 향상 차트로 표현하였다.



〈Figure 5〉 Decile-Wise Lift Chart



〈Figure 6〉 Lift Chart

성능 평가의 결과를 통해 알 수 있듯이, 제안된 모형을 통해 상품 간 잠재적 연관성을 예측하는 경우, 모형을 사용하지 않고 임의로 미연관 상품 간 연관성을 예측하는 방법에 비해 월등히 높은 성과를 얻을 수 있다. 이는 제안 모형을 활용함으로써, 아직 드러나지 않은 상품 간의 연관성을 효과적으로 찾아 소비자에게 제시할 수 있음을 시사한다. 특히, 구매 건수가 충분히 확보되지 않은 신규 상품에 대해 효과적인 추천 목록을 생성할 것이며, 이로 인한 매출 증가 효과를 기대할 수 있을 것이다.

5. 결론

본 연구는 낮은 거래 빈도로 인해 잘 드러나지 않는 상품 간의 잠재적인 연관성을 찾아 연관 규칙을 확장하기 위한 목적으로 수행되었다. 상품 간 연관성에 기반한 연관 상품 추천은 온라인 상거래에서 소비자의 상품 탐색 시간을 줄여줄 뿐만 아니라 판매자의 매출을 증대하는 데에도 크게 기여하고 있다. 그러나, 연관 상품을 추천하는 근거가 되는 연관 규칙은 주문과 같은 거래 건수의 빈도를 기반으로 생성되므로, 신규 상품과 같이 초기에 충분한 거래 건수가 쌓이지 않는 상품은 다른 상품과 연관 상품으로 연결되기 어려운 콜드 스타트(Cold Start) 문제가 제기된다. 연관 상품 추천에서 누락된 상품은 소비자에게 노출될 기회를 잃어 상대적으로 거래 건수를 확보할 기회를 잃게 되는 악순환을 겪을 수도 있다. 이와 같이 잠재된 연관성은 거래가 지속되고 해당 상품들이 함께 출현하는 거래의 건수가 늘어 연관 규칙에 포함될 정도의 임계치를 넘게 되면 많은 경우 자연스럽게 드러날 수 있을 것이

다. 그러나, 이런 임계치의 거래 건수에 미치는 기간이 길어지면 그 기간 동안 해당 상품의 거래는 정체될 수밖에 없다. 예를 들어, 의류 등과 같이 유행에 민감하거나 계절 변화에 영향을 많이 받는 상품의 경우에는 상품 출시 초기에 소비자에게 노출되는지의 여부가 매출에 매우 큰 영향을 미칠 수 있다는 점에서 이런 잠재적 연관성을 미리 발견하고 상품의 노출기회를 확보하는 것이 필요하다.

두 상품 간에 직접적인 연관성이 발견되지 않는다 하더라도 다른 상품을 매개로 두 상품 간에 간접적인 연관성이 존재한다면 이를 활용하여 두 상품 간의 잠재적 연관성을 예측할 수 있을 것이며, 이런 연관성은 여러 상품 간에 서로 영향을 미치는 상호작용의 형태로 나타날 것이므로 사회연결망 분석기법을 활용한 분석방법을 시도하였다. 3장의 연구 모형에서 보인 것처럼 미연관 상품 간 연관성은 두 상품 간을 잇는 경로의 특성과 각 상품이 사회연결망에서 갖는 특성에 영향을 받는다는 것을 보였다. 즉, 두 상품 간 경로의 최단 거리 및 경로의 개수, 각 상품이 얼마나 많은 상품과 연관성을 갖는지, 두 상품의 분류 카테고리가 어느 정도 일치하는지가 두 상품 간의 연관성에 영향을 미친다. 이 모형으로부터 미연관 상품 간 연관성을 예측할 수 있을 것으로 기대되어, 4장에서는 모형의 성능을 평가하고자 실험을 실시했다. 구체적으로는, 전체 거래 중 마지막 10일 간의 주문 거래를 제외한 채로 상품 간 주문 연결망을 구성하고, 이로부터 제외된 10일 간 생성될 상품 간 연관성을 예측하는 방법으로 실험을 진행하였다. 실험 결과를 통해, 모형을 적용하지 않고 찾을 수 있는 연관성의 수에 비해 제안 모형은 훨씬 많은 수의 연관성을 찾을 수 있음을 확인할 수 있었다.

본 연구는 노출 시기가 중요한 상품의 경우 유용하게 활용될 수 있을 것으로 기대된다. 특히, 유행이나 계절 등의 영향을 많이 받는 상품이거나, 스마트폰 앱 등과 같이 신상품의 출시가 빠르게 일어나며 그 수명주기가 짧은 상품일수록 더 큰 효과를 보일 것으로 기대된다. 또한, 의료 분야에서 발병 빈도가 낮아 진단하기 힘든 희귀병을 조기에 진단하는 데에도 활용할 수 있을 것이다. 사회연결망 분석은 연결망을 구성하는 노드와 링크의 수에 매우 민감하게 복잡도가 높아지는 특성을 갖고 있기 때문에 전체 주문을 분석 대상으로 다루는 것은 현실적인 한계를 갖는다. 이런 이유로, 본 연구는 잡화라는 특정한 상품 분류 카테고리에 국한되어 수행되어 일반화의 한계를 가질 수 있다. 이는 서로 다른 분류 카테고리에 속한 상품 간에 존재할 수 있는 의외의 연관성을 발견하는 기회를 제약할 수 있다.

참고문헌(References)

Agrawal, R., T. Imielinski, A. Swami. “Mining association rule between sets of items in large databases,” *Proc. 1993 ACM SIGMOD international conference on management of data*, (1993), 207~216.

Adomavicius, G., A. Tuzhilin. “Context-Aware Recommender Systems. *Recommender Systems Handbook*, Springer US, (2011), 217~253.

Anand, S.S., A.R. Patrick. “A Data Mining methodology for cross-sales,” *Knowledge-Based Systems*, Vol.10**,** No.7(1998), 449~461.

Ansari, A., S. Essegaier, R. Kohli. “Internet recommender systems,” *Journal of Marketing Research*, Vol.37, No.3(2000), 363~375.

Balabanovic, M., Y. Shoham. "Content-Based, Collaborative, Recommendation," *Communications of the ACM*, Vol.40, No.3 (1997), 66~72.

Bodapati, A.V. "Recommender systems with purchase data," *Journal of Marketing Research,* Vol.45, No.1(2008), 77~93.

Chen, Y.L., J.M. Chen, C.W. Tung. "A data mining approach for retail knowledge discovery with consideration of the effect of shelf-space adjacency on sales," *Decision Support Systems,* Vol.42, No.3(2006), 1503~1520.

Choi, S., Hyun, Y., Kim, N. "Improving Performance of Recommendation Systems Using Topic Modeling," *Journal of Intelligence and Information Systems*, Vol.21, No.3(2015), 101~116.

Choi, S., Kwahk, K.-Y., Ahn, H. "Enhancing Predictive Accuracy of Collaborative Filtering Algorithms using the Network Analysis of Trust Relationship among Users," *Journal of Intelligence and Information Systems*, Vol.22, No.3(2016), 113~127.

Fleder, D., K. Hosanagar. "Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity," *Management Science,* Vol.55, No.5(2009), 697~712.

Kang, B. S., "A Novel Web Recommendation Method for New Customers Using Structural Holes in Social Networks," Journal of Industrial Economics and Business, Vol.23, No.5(2010), 2371~2385.

Kim, H. K., Choi, I. Y., Ha, K. M., Kim, J. K. "Development of User Based Recommender System using Social Network for u-Healthcare," *Journal of Intelligence and Information Systems*, Vol.16. No.3(2010), 181~199.

Kim, B. K., S. Lee, S. Bang, J. Kim, and J. H. Lee, "Personalized Recommendation System Using Social Network," Proceedings of the Conference on Intelligent Information Systems, Vol.20, No.1(2010), 48~49.

Kim, J., Lee, S.-W. "The Ontology Based, the Movie Contents Recommendation Scheme, Using Relations of Movie Metadata," *Journal of Intelligence and Information Systems*, Vol.19, No.3(2013), 25~44.

Kim, K.-J., Kim, B.-G. "Product Recommender System for Online Shopping Malls using Data Mining Techniques," *Journal of Intelligence and Information Systems*, Vol.11, No.1(2005), 191~205.

Kim, M., and K. J. Kim, "Recommender Systems using Structural Hole and Collaborative Filtering," *Journal of Intelligence and Information Systems*, Vol.20, No.4(2014), 107~120.

Kim, M. G., and K. J. Kim, " Recommender Systems using SVD with Social Network Information," *Journal of Intelligence and Information Systems*, Vol.22, No.4(2016), 1~18.

Kim, S. H., and R. S. Chang, "The Study on the Research Trend of Social Network Analysis and the its Applicability to Information Science," *Journal of the Korean Society for Information Management*, Vol.27, No.4 (2010), 71~87.

Kim, Y., W.N. Street. "An intelligent system for customer targeting: a data mining approach," *Decision Support Systems,* Vol.37, No.2 (2004), 215~228.

Konstan, J.A., B.N. Miller, D. Maltz, J.L. Herlocker, L.R. Gordon, J. Riedl. "GroupLens: applying collaborative filtering to Usenet news," *Communications of the ACM*, Vol.40, No.3(1997), 77~87.

Lee, D., S. Park, S. Moon. "Utility-based association rule mining: A marketing solution for cross-selling," *Expert Systems with Applications*. Vol.40, No.7(2013), 2715~25.

Noh, H., S. Choi, and H. Ahn, "Social Network-based Hybrid Collaborative Filtering using Genetic Algorithms," *Journal of Intelligence and Information Systems*, Vol.23, No.2(2017), 19~38.

Park, J. H., Y. H. Cho, and J. K. Kim, "Social Network : A Novel Approach to New Customer Recommendations," *Journal of Intelligence and Information Systems*, Vol.15, No.1(2009), 123~140.

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.

Yun, Y., and S. Chae, *Introduction to Complex Systems*, Samsung Economic Research Institute, 2005.

Sohn D., *Social Network Analysis*, Kyungmoon Publications, 2002.

Y. Kim, *Social Network Analysis*, Pakyoungsa, 2003.

Abstract

Extension Method of Association Rules Using Social Network Analysis

Dongwon Lee*

Recommender systems based on association rule mining significantly contribute to seller's sales by reducing consumers' time to search for products that they want. Recommendations based on the frequency of transactions such as orders can effectively screen out the products that are statistically marketable among multiple products. A product with a high possibility of sales, however, can be omitted from the recommendation if it records insufficient number of transactions at the beginning of the sale. Products missing from the associated recommendations may lose the chance of exposure to consumers, which leads to a decline in the number of transactions. In turn, diminished transactions may create a vicious circle of lost opportunity to be recommended. Thus, initial sales are likely to remain stagnant for a certain period of time. Products that are susceptible to fashion or seasonality, such as clothing, may be greatly affected.

This study was aimed at expanding association rules to include into the list of recommendations those products whose initial trading frequency of transactions is low despite the possibility of high sales. The particular purpose is to predict the strength of the direct connection of two unconnected items through the properties of the paths located between them. An association between two items revealed in transactions can be interpreted as the interaction between them, which can be expressed as a link in a social network whose nodes are items. The first step calculates the centralities of the nodes in the middle of the paths that indirectly connect the two nodes without direct connection. The next step identifies the number of the paths and the shortest among them. These extracts are used as independent variables in the regression analysis to predict future connection strength between the nodes. The strength of the connection between the two nodes of the model, which is defined by the number of nodes between the two nodes, is measured after a certain period of time. The regression analysis results confirm that the number of paths between the two products, the distance of the shortest path, and the number of neighboring items connected to the products are significantly related to their potential strength.

* Corresponding Author: Dongwon Lee
School of Business Administration, College of Social Sciences, Hansung University
116 Samseongyoro-16gil, Seongbuk-gu, Seoul 02876, Korea
Tel: +82-2-760-4250, Fax: +82-2-760-4482, E-mail: dongwonlee@hansung.ac.kr

This study used actual order transaction data collected for three months from February to April in 2016 from an online commerce company. To reduce the complexity of analytics as the scale of the network grows, the analysis was performed only on miscellaneous goods. Two consecutively purchased items were chosen from each customer's transactions to obtain a pair of antecedent and consequent, which secures a link needed for constituting a social network. The direction of the link was determined in the order in which the goods were purchased. Except for the last ten days of the data collection period, the social network of associated items was built for the extraction of independent variables. The model predicts the number of links to be connected in the next ten days from the explanatory variables. Of the 5,711 previously unconnected links, 611 were newly connected for the last ten days. Through experiments, the proposed model demonstrated excellent predictions. Of the 571 links that the proposed model predicts, 269 were confirmed to have been connected. This is 4.4 times more than the average of 61, which can be found without any prediction model.

This study is expected to be useful regarding industries whose new products launch quickly with short life cycles, since their exposure time is critical. Also, it can be used to detect diseases that are rarely found in the early stages of medical treatment because of the low incidence of outbreaks. Since the complexity of the social networking analysis is sensitive to the number of nodes and links that make up the network, this study was conducted in a particular category of miscellaneous goods. Future research should consider that this condition may limit the opportunity to detect unexpected associations between products belonging to different categories of classification.

Key Words : Recommendation system, Association rule mining, Social network analysis, Association rule extension, Cold start problem

Received : July 31, 2017 Revised : September 17, 2017 Accepted : September 20, 2017
Publication Type : Regular Paper Corresponding Author : Dongwon Lee

저 자 소 개

이 동 원
LG CNS에서 시스템 엔지니어로 근무하였으며, KAIST 경영대학원에서 MIS 전공으로 석사/박사 학위를 취득하였다. 현재 한성대학교 경영학부 조교수로 재직 중이다. 현재 빅데이터에 기반한 연구를 주로 수행하고 있으며, 주요 관심분야는 고객관계관리, 추천 시스템, 데이터 마이닝 기법의 정교화, 디지털 콘텐츠 마케팅 등이다.