Journal of the Korean Institute of Industrial Engineers
Vol. 42, No. 4, pp. 249-255, August 2016.
ISSN 1225-0988 | EISSN 2234-6457
http://dx.doi.org/10.7232/JKIIE.2016.42.4.249

<Original Research Paper>

트랜잭션 데이터 분석을 위한 확률 그래프 모형

안길승 · 허 선†

한양대학교 산업경영공학과

Probabilistic Graphical Model for Transaction Data Analysis

Gil Seung Ahn · Sun Hur

Department of Industrial and Management Engineering, Hanyang University

Recently, transaction data is accumulated everywhere very rapidly. Association analysis methods are usually applied to analyze transaction data, but the methods have several problems. For example, these methods can only consider one-way relations among items and cannot reflect domain knowledge into analysis process. In order to overcome defect of association analysis methods, we suggest a transaction data analysis method based on probabilistic graphical model (PGM) in this study. The method we suggest has several advantages as compared with association analysis methods. For example, this method has a high flexibility, and can give a solution to various probability problems regarding the transaction data with relationships among items.

Keywords: Probabilistic Graphical Model, Transaction Data, Association Rule, Point-Wise Mutual Information

1. 서 론

트랜잭션 데이터(transaction data)란 은행에서의 고객의 입출금, 상점에서의 고객의 주문, 웹에서의 사용자의 클릭 등을 기록한 데이터를 말한다. 스마트폰의 활성화와 핀테크의 출현 등으로 세계 전자 상거래의 시장이 꾸준히 커가고 있으며(전자 상거래를 이용한 세계 B2C 거래액은 2013년 1조 2천억 달러, 2014년 1조 5천억 달러이며 2018년에는 2조 4천억 달러까지 증가할 전망(eMarketer, 2014)), 이로 인해 트랜잭션 데이터는 엄청난 속도로 쌓여갈 것이라 기대된다.

트랜잭션 데이터는 고객의 행위와 직접 관련되어 있어, 고객 세분화와 구매 여부 예측 등 고객을 이해하고 효과적인 마케팅 활동을 수행하는 데에 필수적이다. 예를 들어, 미국의 소비재 기업과 유통업체를 고객사로 하는 마케팅 전문업체 C사는 여러 소매업체로부터 매주 2억 5천만 건 이상의 트랜잭션 데이터를 수집해서 구매자의 구매 행동을 분석하고 예측하여 개인화된 맞춤 쿠폰 및 정보지를 제작하고 배포함으로써 엄청난 매출을 올렸다(Informationweeek, 2012).

연관분석(association mining)은 하나의 트랜잭션에 포함된 아이템의 관련성을 파악하여 둘 이상의 아이템으로 구성된 연관규칙을 도출하는 탐색적 자료 분석 방법으로, 트랜잭션 데이터를 분석하는 데 가장 흔히 사용되는 방법론이다. 그러나 연관분석을 이용한 트랜잭션 데이터 분석에는 몇 가지 문제가 있다. 예를 들어, 연관분석은 탐색적 자료 분석 방법이므로 수리적인 모형이 존재하지 않아 일반화가 불가능하다. 또한, 특정 분야의 도메인 지식(domain knowledge)을 반영하기 어려우며, 연관분석의 결과로 생성되는 규칙을 일관되게 평가할 만한 기준이 없다.

확률 그래프 모형(probabilistic graphical model)은 복잡한 구조를 가진 데이터에 포함된 변수 간의 조건부 관계(conditional relationship)를 표현하는 확률 모형이다. 확률 그래프 모형은 변수가 많고 변수 간에 상관관계가 있는 복잡한 구조의 데이터를 분석하기에 적합하다. 또한, 확률 그래프 모형은 결측 데이터가 다수 존재하는 경우에도 우수한 성능을 나타내며 높은 설명력을 가지는 등의 장점이 있다(Jordan, 1999). 이에 따라 확률 그래프 모형은 복잡한 구조의 데이터가 대량으로 생산되는 자연어 처리, 음성 인식, 컴퓨터 비전(computer vision) 등의 분야에서 활발히 사용되고 있다. 특히, 고객과 고객, 고객과 상품,

† 연락저자 : 허 선 교수, 15588 경기도 안산시 상록구 한양대학로 55, 한양대학교 산업경영공학과, Tel : 031-400-5265, Fax : 031-400-5265, E-mail: hursun@hanyang.ac.kr
2015년 12월 7일 접수, 2016년 5월 1일 수정본 접수, 2016년 5월 27일 게재 확정.

상품과 상품 간 관계가 매우 다양한 CRM(Customer Relationship Management) 상황에서도 확률 그래프 모형이 활용되고 있다. 예를 들어, Ahn and Hur(2015)의 연구에서는 확률 그래프 모형의 일종인 continuous conditional random field를 이용하여 고객간의 관계를 분석함으로써 신규 고객의 충성도를 예측하는 방법을 제시하였다.

확률 그래프 모형을 다음과 같은 이유로 연관분석 대신 트랜잭션 데이터 분석에 적용할 수 있다. 첫째, 도메인이 같다면 일반화가 가능하며, 이러한 점을 이용하여 예측이나 분류를 수행할 수 있다. 둘째, 결합확률분포를 제공하기 때문에 이를 변형하면 특정 아이템의 등장 확률과 조건부 확률 등을 계산하는데 응용할 수 있다. 셋째, 포텐셜 함수(potential function)를 통해 도메인 지식을 반영할 수 있다. 넷째, 각 포텐셜 함수에 부여된 가중치를 통해 일관되게 규칙을 평가할 수 있다.

이에 본 연구에서는 연관분석의 문제점을 보완할 수 있도록 대표적인 확률 그래프 모형인 마코프 네트워크(Markov network) 기반의 트랜잭션 분석 방법론을 제안한다. 그리고 본 연구에서 제안하는 방법론의 활용 방안과 성능 평가 방안에 대해 제시한다. 특히, 성능 평가는 비지도 학습(unsupervised learning)인 연관분석에서 불가능한 내용이다.

본 논문은 다음과 같이 구성되어 있다. 제 2장에서는 연관분석, 마코프 네트워크, point-wise mutual information(PMI)의 개념을 설명하고 관련 연구를 소개한다. 제 3장에서는 본 연구에서 제안하는 알고리즘을 크게 다섯 단계로 구분하여 설명한다. 또한, 제안 알고리즘의 다양한 활용 방안에 관해 설명한다. 제 4장에서는 본 연구에서 제안하는 방법의 성능을 평가하기 위해 실제 트랜잭션 데이터에 적용하는 과정을 설명한다. 마지막으로 제 5장에서는 본 연구를 정리한다.

2. 관련 연구

2.1 연관분석

연관분석은 데이터베이스 내에서 빈발하게 발생하는 패턴을 추출하고, 트랜잭션 내 아이템 사이의 연관성을 파악하는 비지도 학습 알고리즘으로 CRM 등의 분야에서 활발하게 사용되고 있다. 일반적으로 연관분석은 사용자가 설정한 최소지지도(minimum support)를 넘는 빈발항목으로 구성된 빈발항목 집합을 만든 후, 빈발항목 집합 내에서 최소 신뢰도(minimum confidence)를 넘는 항목을 찾아내는 순서로 수행한다. 여기서 지지도란 아이템 i와 아이템 j가 동시에 포함된 트랜잭션의 비율을 말하며, 신뢰도란 아이템 i를 포함하는 트랜잭션 중 아이템 j를 포함하고 있는 트랜잭션의 비율을 의미한다. 즉, 지지도와 신뢰도는 다음과 같이 계산한다.

$$\text{Support}(i \rightarrow j) = \frac{\text{아이템 } i\text{와 } j\text{를 포함한 트랜잭션 수}}{\text{전체 트랜잭션 수}}, \quad (1)$$

$$\text{Confidence}(i \rightarrow j) = \frac{\text{아이템 } i\text{와 } j\text{를 포함한 트랜잭션 수}}{\text{아이템 } i\text{를 포함한 트랜잭션 수}}. \quad (2)$$

즉, 연관성 분석의 핵심은 고객이 상품 A를 구매할 경우 B를 구매할 조건부 확률을 계산하는 것이다(Son *et al.*, 2015).

연관분석을 트랜잭션 데이터를 분석하는데 적용한 연구는 크게 상품 추천 시스템 설계, 계산량 등을 비롯한 연관분석의 내재적 한계를 극복하기 위한 연구, 기존 평가 척도의 한계를 극복하기 위한 새로운 척도를 개발하기 위한 연구로 구분할 수 있다.

상품 추천 시스템 설계와 관련된 연구로, 연관규칙을 이용하여 웹상에서의 상품 추천 시스템을 구현한 연구(Park, 2005)를 들 수 있다. 그러나 이 연구에서는 트랜잭션 내 출현횟수가 적은 아이템에 관해서는 추천을 수행하지 못했다는 한계가 있다.

연관분석의 내재적 한계를 극복하기 위한 연구로, Kim(2008)과 Yang(2003)을 들 수 있다. Kim(2008)에서는 크기가 큰 트랜잭션에 포함된 아이템일수록 지지도와 신뢰도가 높게 나오는 현상을 지적하면서, 트랜잭션의 크기에 반비례한 가중치를 아이템에 부여하여 이러한 문제를 해결하였다. 한편, Yang(2003)에서는 아이템의 수량적 속성을 고려하기 위해, 긴밀성을 나타낼 수 있는 긴밀 계수를 정의하였다. 여기서 긴밀성이 높을수록 두 아이템 간의 상호 의존도가 높아서 구매 상승효과를 기대할 수 있다.

연관분석의 새로운 척도를 개발하기 위한 연구 역시 다수 이루어졌다. 예를 들어, Han(2009)에서는 단백질의 기본 단위 체인 도메인의 조합을 찾아내는데 연관분석을 적용하였다. 이 연구에서는 연관분석은 단방향의 규칙만 생성하기 때문에, 도메인 조합이 제한적일 수밖에 없다는 사실을 지적하였다. 그들은 이러한 문제를 해결하기 위해 all-confidence를 정의함으로써 양방향성의 규칙을 생성할 수 있는 연관분석을 수행하였다. 한편, Fuguang(2015)에서는 기존 연관규칙을 평가하기 위한 객관적인 척도가 없고, 통계적 기반이 부족하고, 아이템 간, 음의 관계는 설명할 수 없다는 점을 지적하면서, bi-lift, bi-improve, bi-confidence라는 새로운 척도를 제안했다.

2.2 마코프 네트워크

마코프 네트워크는 변수들의 결합확률분포를 네트워크 형태로 표현한 모형으로, 비방향성의 그래프(undirected graph) G와 변수 간의 관계를 나타내는 포텐셜 함수 ϕ_k로 구성된다. 여기서 그래프 G에 속한 노드(node) i는 확률벡터 $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)$에서 확률변수 X_i의 값을 나타내고, 노드 i와 노드 j를 연결하는 에지(edge)는 확률변수 X_i와 X_j간에 어떠한 관계가 존재함을 의미한다. 또한, ϕ_k에서 색인 k는 네트워크 내에 존재하는 모든 노드가 완전히 연결된 부분 그래프인 클리크(clique) k를 나타낸다. 즉, 포텐셜 함수는 각 클리크에 포

함된 변수 간의 관계를 표현한다. 이처럼, 한 클리크에 속한 노드들은 서로 의존적이나 다른 클리크에 속한 노드들은 서로 독립적이다. 따라서 마코프 네트워크는 다음과 같은 조건을 만족한다(Rajtmajer, 2012).

$$P(\boldsymbol{X}=\boldsymbol{x}) > 0, \text{ for all } \boldsymbol{x}, \tag{3}$$

$$P(X_i = x_i|\boldsymbol{X}_{(-i)} = \boldsymbol{x}_{(-i)}) = P(X_i = x_i|\boldsymbol{X}_{N_i} = \boldsymbol{x}_{N_i}), \tag{4}$$

위 식에서 $\boldsymbol{X}_{(-i)}$와 $\boldsymbol{X}_{N_i}$는 각각 노드 i를 제외한 모든 노드, 노드 i와 에지로 연결된 모든 노드를 의미한다.

클리크의 특성을 이용하여 확률벡터 $\boldsymbol{X}$에 대한 결합확률분포는 다음과 같이 클리크의 포텐셜 함수의 곱으로 표현할 수 있다(Kolaczyk, 2009).

$$P(\boldsymbol{X}=\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \Pi_k \phi_k(\boldsymbol{x}_{\{k\}}), \tag{5}$$

위 식에서 $\boldsymbol{x}_{\{k\}}$는 클리크의 상태(configuration)를 나타내는 값이며, 포텐셜 함수는 사용자가 정의한다. 또한, $Z(\boldsymbol{x})$는 $\Pi_k \phi_k(\boldsymbol{x}_{\{k\}})$의 값을 0과 1사이의 값으로 바꿔주기 위한 정규화 함수(normalization function)로, 다음과 같이 계산한다.

$$Z(\boldsymbol{x}) = \Sigma_{\boldsymbol{x}\in\boldsymbol{X}} \Pi_k \phi_k(\boldsymbol{x}_{\{k\}}). \tag{6}$$

계산의 편의를 위해, 마코프 네트워크를 일반적으로 선형 로그모형(log-linear model)으로 다음과 같이 변환하여 사용한다(Chen and Welling, 2014).

$$P(\boldsymbol{X}=\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \exp(\Sigma_k \omega_k f_k(\boldsymbol{x}_{[k]})). \tag{7}$$

위 식에서 ω_k와 f_k는 각각 클리크 k에 부여된 가중치와 클리크 k의 상태를 나타내는 실함수(real-valued function)로, 특징함수(feature function)라고 부른다. $\boldsymbol{x}_{[k]}$는 클리크 k를 구성하는 노드들의 벡터이다. 식 (7)에 제시된 선형 로그모형은 식 (3)과 식 (4)에 나타난 마코프 네트워크의 두 가지 조건을 만족한다. 구체적으로, $\exp(\Sigma_k \omega_k f_k(\boldsymbol{x}_{[k]}))$은 $\Sigma_k \omega_k f_k(\boldsymbol{x}_{[k]})$값과 관계없이 항상 양수이므로 식 (3)을 만족하고, 클리크 $\boldsymbol{x}_{[k]}$를 기준으로 특징함수가 구성되므로 임의의 X_i의 값은 그 이웃인 $\boldsymbol{X}_{N_i}$의 값에만 영향을 받는다. 즉, 식 (4)를 만족한다.

마코프 네트워크의 파라미터인 가중치 ω_k의 학습을 효율적으로 수행하기 위한 여러 연구 가운데 가장 보편적으로 쓰이는 방법은 최대 로그 가능도 방법(maximum log likelihood method)이다. 즉, M개의 레코드가 있을 때, 식 (7)의 파라미터 집합 $\boldsymbol{\omega}$를 추정하기 위해서는 다음과 같은 식을 최대화하는 $\boldsymbol{\omega}$를 찾아야 한다.

$$\log l(\boldsymbol{\omega}) = \Sigma_M \Sigma_{k\in C} \log f_k(\boldsymbol{x}_{[k]}) - M \log Z(\boldsymbol{\omega}), \tag{8}$$

여기서 C는 클리크의 집합이다. $\boldsymbol{\omega}$의 추정값 $\hat{\boldsymbol{\omega}}$는 다음을 만족한다.

$$\hat{\boldsymbol{\omega}} = \arg\max_{\boldsymbol{\omega}} (\log l(\boldsymbol{\omega})). \tag{9}$$

한편, 추론(inference)이란 $\hat{\boldsymbol{\omega}}$가 주어졌을 때, $P(\boldsymbol{X}=\boldsymbol{x})$를 최대화하는 $\hat{\boldsymbol{x}}$를 찾는 작업이다 :

$$\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} P(\boldsymbol{X}=\boldsymbol{x}|\hat{\boldsymbol{\omega}}). \tag{10}$$

2.3 Point-wise Mutual Information

Point-wise mutual information(PMI)는 텍스트마이닝(text mining)에서 두 단어 간의 연관성을 표현하는 지표로, 두 단어 w_1과 w_2의 PMI는 다음과 같이 계산한다(Turney, 2002).

$$\text{PMI}(w_1, w_2) = \log_2(\frac{p(w_1, w_2)}{p(w_1)p(w_2)}), \tag{11}$$

위 식에서 $p(w_j)(j=1, 2)$는 단어 w_j가 출현한 문서의 비율을, $p(w_1, w_2)$는 w_1과 w_2가 동시에 출현한 문서의 비율을 나타낸다. 만약 w_1과 w_2가 연관성이 전혀 없다면, 즉 독립이라면, $p(w_1, w_2)$는 0이므로 식 (11)은 음의 무한대로 발산한다. 한편, 만약 w_1과 w_2가 완전히 연관되어 있다면 $p(w_1, w_2) = \min(p(w_1), p(w_2))$가 성립하므로,

$$\begin{aligned}\text{PMI}(w_1, w_2) &= \log_2(\frac{\min(p(w_1), p(w_2))}{p(w_1)p(w_2)}) \\ &= \log_2(\frac{1}{\max(p(w_1), p(w_2))}) \\ &= -\log_2(\max(p(w_1), p(w_2)))\end{aligned}$$

역시 성립한다. 다시 말해, $\text{PMI}(w_1, w_2)$의 값의 범위는 $(-\infty, \max(-\log_2 p(w_1), -\log_2 p(w_2))$이 되며, 이 값이 클수록 단어 간의 연관성이 높으며, 음의 무한대에 가까우면 연관성이 낮다고 할 수 있다.

이러한 PMI의 개념을 트랜잭션 데이터에 포함된 아이템 간의 유사도를 계산하는데 적용할 수 있다. 유사도를 구하기 위한 두 아이템을 i_1과 i_2라 하면, $p(i_j)$은 아이템 $i_j(j=1, 2)$가 각 트랜잭션에 출현한 비율을 나타내며 $p(i_1, i_2)$는 아이템 i_1과 i_2가 각 트랜잭션에 동시에 출현한 비율을 나타낸다. 즉, 아이템 i_1과 i_2를 같이 구매한 비율을 나타낸다. 본 연구에서는 이러한 개념을 아이템 간의 유사도를 계산하는 데 활용한다.

3. 제안 알고리즘

본 연구에서 제안하는 알고리즘을 도식화하면 <Figure 1>과 같다.

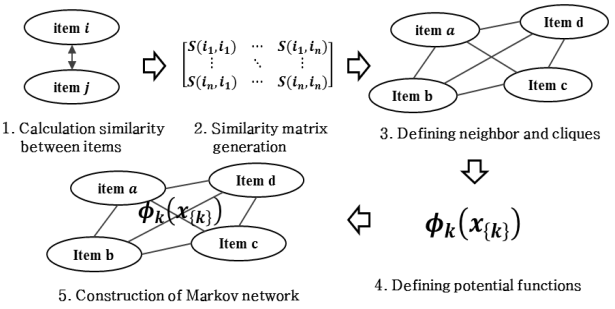

Figure 1. Suggested Method

제안 알고리즘의 첫 단계로 아이템 간 유사도를 계산하며, 아이템 i와 아이템 j간의 유사도는 다음과 같이 계산한다.

$$\mathrm{Sim}(item_i, item_j) = \log_2\left(\frac{p(item_i, item_j)}{p(item_i)p(item_j)}+1\right), \quad (12)$$

여기서 $p(item_i)$와 $p(item_i, item_j)$는 각각 아이템 i가 트랜잭션에 출현한 비율과 아이템 i와 j가 동시에 트랜잭션에 출현한 비율을 나타낸다. 또한, 기존의 PMI 식과 다르게 로그의 진수에 1을 더한 것은 로그의 진수가 0이 되는 것을 방지하기 위함이다. 또한, 식 (12)의 범위는 $[0, \log_2(\min(p(item_i), p(item_j)))]$ 로 일반적인 유사도 범위인 [0, 1]이 아니다. 따라서, 식 (12)가 최댓값을 가질 수 있는 경우인 $\max_i \mathrm{Sim}(item_i, item_i)$로 식 (12)를 나눠줌으로써 범위를 [0, 1]로 스케일링(scaling)하였다. 이는 가장 많이 등장한 아이템 $item_i$에 대해, $\log_2\left(\frac{1}{p(item_i)}+1\right)$을 계산한 것과 같다.

유사도 행렬 S는 식 (12)에서 제시한 아이템 간 유사도를 바탕으로 구성하며, 그 방법은 다음과 같다.

$$S_{ij} = \mathrm{Sim}(item_i, item_j). \quad (13)$$

유사도 행렬에서 값이 임계치 α 이상이면 두 아이템을 이웃이라 정의한다. 여기서 α는 사용자 정의 파라미터(user parameter)이다.

이제, 정의된 이웃을 바탕으로 최대클리크(maximal clique)를 파악해야 하는데, 여기서 최대클리크란 더는 다른 노드를 클리크에 추가할 수 없는 클리크를 말한다. 노드 수가 적을 때는 최대클리크를 찾는 작업은 어렵지 않으나 노드가 많을 때는 그렇지 않다. 이러한 최대클리크 문제를 해결하기 위해 사용되는 최대클리크 알고리즘과 관련된 연구가 다수 이루어졌으며, 이와 관련된 대표적인 알고리즘으로 Bron-Kerbosch 알고리즘을 들 수 있다(Bron and Kerbosch, 1973).

최대클리크를 모두 파악한 후, 각 최대클리크에 대응되는 포텐셜 함수를 정의해야 한다. 포텐셜 함수는 사용자가 각 도메인에 맞게 설정할 수 있다. 예를 들어, 이미지 프로세싱(image processing)에서 픽셀(pixel)은 노드를 의미하는데, 가까운 픽셀 간 색상은 유사할 확률이 높으므로 클리크에 포함된 노드 값이 같으면 1점을, 그렇지 않으면 -1점을 부여하는 식이다. 또한, 트랜잭션 데이터에 맞도록 포텐셜 함수를, 같은 트랜잭션에 포함될 가능성이 큰 아이템이 포함되어 있으면 1점을 그렇지 않으면 -1점을 부여하는 식으로 정의할 수 있다. 각 점수는 포텐셜 함수에 부여되는 각각의 가중치에 의해 자동으로 조정되므로, 0을 제외한 어떠한 점수를 부여해도 상관없다.

식 (9)를 이용하여, 이들 포텐셜 함수에 부여된 가중치를 학습함으로써 마코프 네트워크를 완성한다. 이렇게 완성한 마코프 네트워크는 트랜잭션 데이터 분석에 다양하게 활용할 수 있다. 예를 들어, 마코프 네트워크의 기본 형태인 $P(\boldsymbol{X} = \boldsymbol{x})$를 이용하여, 특정 아이템 집합은 트랜잭션에 포함되면서 다른 아이템 집합은 트랜잭션에 포함되지 않을 확률을 계산하는 데 사용할 수 있다. 다른 예로, 특정 아이템이 트랜잭션에 등장하였을 때 다른 아이템이 트랜잭션에 등장할 확률을 계산하는 데 사용할 수도 있다. 물론, 특정 아이템이 트랜잭션에 등장할 확률 역시 계산할 수 있다.

4. 제안 알고리즘의 적용 예와 성능 평가

이 장에서는 본 연구에서 제안하는 알고리즘을 예시한다. 또한, 그 성능을 판단하기 위해 실제 데이터에 적용하여 실험하고 그 결과를 평가한다.

4.1 데이터 세트

실험에 사용한 데이터 세트는 'Anonymous Microsoft Web Data'와 'Extend Bakery Dataset'으로 각각 UCI Machine Learning Repository(https://archive.ics.uci.edu/ml/datasets)와 Trac(https://wiki.csc.calpoly.edu/datasets/wiki/apriori)에서 획득하였다. 'Anonymous Microsoft Web Data'는 임의로 선정된 4,183명의 사용자가 일주일간 방문한 사이트 목록이며, 본 연구에 제안한 방법론에 적용하기 위해 데이터의 구조를 <Table 1>과 같이 트랜잭션 데이터 세트 형태로 변경하였다. 또한, 검증 데이터 세트에는 총 5,000명의 사용자가 일주일간 방문한 목록을 포함하고 있다.

Table 1. Anonymous Microsoft Web Data Structure

	Site000	Site001	…	Site233
ID0001	1	1	…	0
ID0002	0	1	…	0
ID0003	0	1	…	0
…	…	…	…	…
ID4183	0	1	…	0

'Extended Bakery Dataset'은 특정 제과점의 1,000명의 고객이 구매한 빵이나 과자의 내역을 포함하고 있는 데이터로, 데이터의 구조는 <Table 2>에 제시하였다.

Table 2. Extended Bakery Data Structure

	Item 1	Item 2	⋯	Item 50
ID0001	0	0	⋯	1
ID0002	0	1	⋯	0
ID0003	0	1	⋯	0
⋯	⋯	⋯	⋯	⋯
ID1000	0	0	⋯	0

원 데이터에는 총 1,000개의 레코드를 포함하고 있으나, 본 연구에서 제안한 방법론의 성능을 객관적으로 평가하기 위해 70%와 30%의 비율로 나누어 학습데이터 세트와 검증데이터 세트를 구성하였다.

4.2 결과

4.2.1 Anonymous Microsoft Web Data

식 (12)와 식 (13)을 이용하여 유사도 행렬을 구성하였는데, 그 결과 유사도 행렬의 성분 대부분이 0인 희소 행렬이었다. 이는 사용자 대부분이 일주일 동안 여러 사이트에 방문하지 않고, 소수의 사이트만 방문하였다는 사실을 나타낸다. 이 행렬 유사도 임계치 α를 0.18로 설정하였을 때, 적당한 수(206개)의 최대 클리크가 생성되었다. 즉, 유사도 행렬의 성분 값의 대다수가 0.18 근처에 있어 α를 0.18보다 작게 설정하면 너무 적은 수의 최대클리크가 생성되었고, 그 반대의 경우에는 너무 많은 수의 최대 클리크가 생성되었다. 구체적으로, α를 0.17로 설정하면 233개의 최대 클리크가 생성되었으며, (즉, 모든 최대 클리크의 크기가 1), α를 0.19로 설정하면 82개의 최대 클리크가 생성되었는데, 이는 유사도가 0인 두 노드를 제외하고는 대부분의 노드가 연결됨을 나타낸다. 따라서 α를 0.18로 설정하여 유사도가 임계치 이상인 두 사이트를 이웃으로 정의한 뒤, 최대클리크를 탐색하였다. 최대클리크는 총 206개였으며, 대다수의 최대 클리크에는 2개 이하의 노드가 포함되어 있었다. 이는 위에서 언급한 바와 같이, 사용자 대부분이 여러 사이트에 방문하지 않고 소수의 사이트만 방문하기 때문에 발생한 현상이라 보인다.

사이트 간 연결을 나타내는 네트워크 구조는 <Figure 2>에 제시하였다. 위에서 언급하였듯이 대부분 사이트는 연결되어 있으나, 세 개 이상의 사이트가 연결된 경우는 드문 것을 확인할 수 있었다.

이제 특징함수를 다음과 같이 정의한다. 최대클리크를 구성하는 모든 노드의 값이 1로 동일하면 1점을, 그렇지 않을 때는 0점을 부여한다. 즉, 노드 집합 $\boldsymbol{x}_{\{k\}}$로 구성된 최대클리크 k에 대응하는 특징함수는 식 (14)와 같다.

$$f_k(\boldsymbol{x}_{\{k\}}) = \begin{cases} 1, & \text{if all components of } \boldsymbol{x}_{\{k\}} \text{ are 1,} \\ 0, & \text{otherwise.} \end{cases} \quad (14)$$

위 특징함수는 이웃으로 정의된 사이트들은 동시에 방문 될 가능성이 크다는 사실을 바탕으로 구성되었다. 또한, 각 사이트의 여러 특징을 고려한 특징 함수 구성도 가능하다. 예를 들어, 포털 사이트가 그렇지 않은 사이트에 비해 방문할 확률이 더 높다는 사실을 바탕으로 특징함수를 구성할 수 있다. 위와 같은 방식으로 은행 입출금 데이터, 장바구니 데이터 등을 비롯한 여러 트랜잭션 데이터 분석에서도 도메인 지식을 특징함수에 반영할 수 있다.

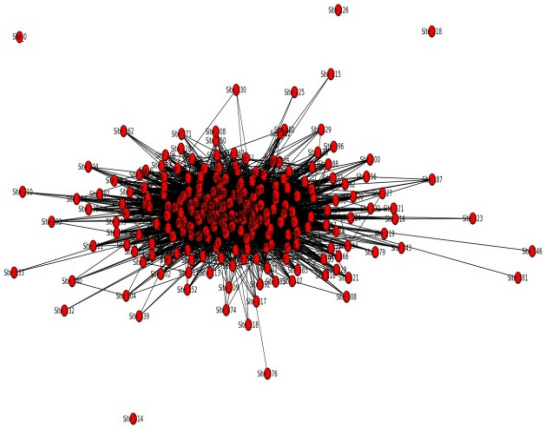

Figure 2. Network Graph of Anonymous Microsoft Web Data

이제 식 (9)를 이용하여 가중치 $\omega_k (k = 1, 2, \cdots, 206)$를 학습한다. 이 때 최적화해야 하는 함수에 포함된 변수가 206개나 되므로 내리막(또는 오르막) 경사법 등을 비롯한 분석적 풀이를 적용하기에는 현실적으로 어려운 점이 많다. 따라서 본 연구에서는 대표적인 메타 휴리스틱 알고리즘인 시뮬레이티드 어닐링(simulated annealing)을 사용하여 가중치를 추정하였으며, 추정한 결과 일부를 <Table 3>에 제시하였다.

Table 3. Learning Results of the First Experiment

Maximal Clique(k)	Nodes($x_1, \cdots, x_t$)	Weight (ω_k)
1	Site003	-6.5214
2	Site006	-4.7813
3	Site007	-3.5332
⋯	⋯	⋯
161	Site016, Site017	8.7372
⋯	⋯	⋯
189	Site014, Site024	-5.2249
⋯	⋯	⋯
206	Site002, Site004, Site005, Site009, Site010, Site018, Site019	1.4195

학습된 모델을 활용하여 다양한 확률값들을 계산할 수 있다. 예를 들어 사용자들이 Site007을 방문했을 확률인 $P(\text{Site007} = 1)$과, 사용자들이 Site16, 17만 방문했을 확률인 $P(\text{Site016} = 1, \text{Site017} = 1, \text{other sites} = 0)$, Site 014를 방문했을 때 Site 024를 방문할 확률인 $P(\text{Site024} = 1 | \text{Site014} = 1)$을 계산해 보자. 우선, $P(\text{Site007} = 1)$은 다음과 같이 계산할 수 있다.

$$P(\text{Site007}=1)=$$
$$\sum_{Site000=0}^{1}\cdots\sum_{Site006=0}^{1}\sum_{Site008=0}^{1}\cdots\sum_{Site233=0}^{1}$$
$$\{\text{P}(\text{Site000, Site001, }\cdots\text{, Site007 = 1, Site008, }\cdots)\}. \quad (15)$$

위 식은 2^{232}번의 연산을 해야 하므로, 현실적으로 계산할 수 없다. 그러나, 식 (7)의 개념을 활용한다면 충분히 계산할 수 있다. 즉, Site007은 <Table 3>에서 보듯이 최대클리크 3(즉 $k=3$)에 포함된 이웃이 없는 독립적인 노드이다. 따라서 다음과 같이 계산할 수 있다.

$$P(\text{Site007}=1)=\exp(\omega_3 f_3(\text{Site007}=1))/Z(\text{Site007}). \quad (16)$$

특징함수 $f_3(\cdot)$은 Site007 = 1일 때만 1이고 그 이외의 경우는 0이다. 다른 모든 특징함수의 값은 0이다. 따라서 식 (16)의 분자의 값은 $\exp(\omega_3 f_3(\text{Site007}=1))=\exp(-3.5332\times 1)$이다. 그리고 분모의 $Z(\text{Site007})$은 식 (17)과 같이 계산할 수 있다.

$$Z(\text{Site007})=\sum_{f_3=0}^{1}\exp(\omega_3 f_3)=1+\exp(-3.53321). \quad (17)$$

결과적으로 식 (16)과 식 (17)을 이용하여 계산한 $P(\text{Site007}=1)$은 0.0284이다.

다음으로 $P(\text{Site016}=1,\text{Site017}=1,\text{other sites}=0)$은 마코프 네트워크가 제공하는 결합확률분포를 이용하면 쉽게 구할 수 있다. <Table 3>에서 보는 바와 같이 Site016과 Site017만 1을 가진다면, 최대 클리크 161번에 대응하는 특징함수 f_{161}만 1점을 가지며 나머지 특징함수는 0점을 가진다. 최대클리크 161번에 대응하는 가중치는 $\omega_{161}=8.7372$이므로 다음과 같이 계산할 수 있다.

$$P(\boldsymbol{X}=\boldsymbol{x})=\frac{\sum_{k=1}^{206}\exp(\omega_k f_k)}{Z(\boldsymbol{x})}$$
$$=\frac{205\times\exp(0)+\exp(8.7372)}{Z(\boldsymbol{x})}=0.0021. \quad (18)$$

마지막으로 Site014를 방문했을 때 Site024를 방문할 확률 $P(\text{Site024}=1\mid\text{Site014}=1)$는 조건부 확률의 기본 식인 $\frac{P(\text{Site024}=1\mid\text{Site014}=1)}{P(\text{Site014}=1)}$을 이용하여 계산할 수 있으며, $P(\text{Site014}=1,\text{Site024}=1)$와 $P(\text{Site014}=1)$는 앞서 설명한 방법과 동일하게 계산할 수 있다. 본 연구에서 제안한 방법론을 이용하여 계산한 결과 그 값은 0.015이다.

본 연구에서 제안한 방법론을 안정성 측면에서 연관분석과 비교하기 위해 Support(Site016 → Site017)와 Confidence(Site014 → Site024)를 다음과 같이 계산한다.

$$\text{Support}(\text{Site016}\rightarrow\text{Site017})=\frac{17}{4183}=0.0041, \quad (19)$$

$$\text{Confidence}(\text{Site014}\rightarrow\text{Site024})=\frac{0}{101}=0.0000. \quad (20)$$

이제 본 연구에서 제안한 방법으로 계산한 값과 연관분석을 통해 식 (19)와 식 (20)에서 계산한 값을 <Table 4>에 비교하여 나타내었다.

Table 4. Calculation Results of Suggested Method and Association Analysis

	P(Site016 = 1, Site017 = 1, other sites = 0)		P(Site024 = 1 | Site014 = 1)	
	Proposed method	Association analysis	Proposed method	Association analysis
Training data set	0.0021	0.0041	0.0153	0.0000
Validation data set	0.0028		0.0162	

계산 결과에서 보듯이, 본 연구에서 제안하는 방법론과 다르게 연관분석의 결과는 데이터에 매우 민감한 것을 알 수 있다. 즉, 학습 데이터 세트에서 계산한 지지도(0.0041) 및 신뢰도(0.0000)와 평가 데이터 세트에서 계산한 지지도(0.0028) 및 신뢰도(0.0162)의 차이가 크다. 이에 반해 본 연구에서 제안한 방법론을 사용하면 학습 데이터 세트에서 계산한 지지도는 0.0021이고 신뢰도는 0.0153으로, 평가데이터 세트에서 계산한 지지도 및 신뢰도와의 차이가 매우 작아서 데이터에 비교적 강건하다고 할 수 있다.

4.2.2 Extended Bakery Dataset

식 (12)와 식 (13)을 이용하여 유사도 행렬을 구성하였으며, 유사도 행렬 성분의 평균이 0.5093이었으며, 0.6보다 큰 성분은 총 424개였다. 이에 유사도 임계치 α를 0.6으로 설정하여 유사도가 임계치 이상인 두 사이트를 이웃으로 정의한 뒤, 최대클리크를 탐색하였다. 탐색결과, 최대클리크는 총 100개였으며, 대부분의 최대클리크 크기는 3이었다. 아이템 간 연결을 나타내는 네트워크 구조는 <Figure 3>에 제시하였다.

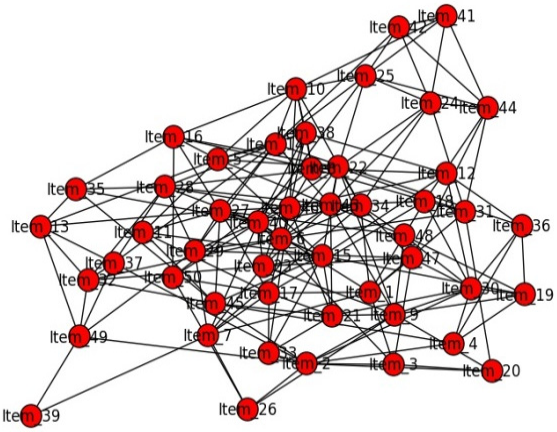

Figure 3. NetworkGraph of Extended Bakery Data

제 4.2.1절의 식 (14)와 동일하게 특징함수를 구성하고 역시 식 (9)를 이용하여 가중치 $\omega_k (k=1, 2, \cdots, 100)$들을 학습하였으며, 학습 알고리즘으로는 시뮬레이티드 어닐링을 사용하였다.

학습된 모델을 바탕으로 사용자들이 Item29와 49만 구매했을 확률인 P(Item29 = 1, Item49 = 1, the other items = 0)을 계산한다. 이는 마코프 모델이 제공하는 결합확률 분포를 이용하여, 식 (18)과 같은 방법으로 다음과 같이 계산할 수 있다.

$$P(\boldsymbol{X}=\boldsymbol{x}) = \frac{\sum_{k=1}^{100}\exp(\omega_k f_k)}{Z(\boldsymbol{x})} = \frac{99+\exp(0.0804)}{Z(\boldsymbol{x})}. \quad (21)$$
$$= 0.0213.$$

이는 평가 데이터 세트에서 계산한 결과인 0.0208과 그 값이 매우 유사한 것을 확인할 수 있다.

또한, 학습된 모델을 바탕으로 각 특징함수를 객관적으로 비교 평가할 수 있다. 예를 들어, Item1과 Item3이 포함된 최대클리크에 정의된 특징함수의 가중치는 0.5821이며, Item1과 Item4가 포함된 최대클리크에 정의된 특징함수의 가중치는 0.3754이다. 이는 Item1과 Item3이 같은 트랜잭션에 등장할 가능성이 Item1과 Item4가 같은 트랜잭션에 등장할 가능성보다 크다는 것을 암시한다. 만약 연관분석을 이용하여 이러한 규칙을 평가한다면, 여러 평가 기준이 존재하여 통합적인 평가가 불가능할 것이다.

5. 결 론

본 연구에서는 최근 다양한 분야에서 엄청난 속도로 축적되고 있는 트랜잭션 데이터를 분석하기 위해, 마코프 네트워크 기반의 방법론을 제안하였다. 해당 방법론에서 아이템 간의 유사도를 PMI를 이용하여 계산하여 유사도 행렬을 구성하였고, 이를 바탕으로 이웃과 클리크를 정의하였다. 마지막으로 각 클리크에 포텐셜 함수를 정의하고 학습시킴으로써 마코프 네트워크를 완성하였다. 또한, 방법론 제시에서 그치는 것이 아니라 본 연구에서 제안한 모형을 실제로 적용하는 데 도움이 될 수 있도록 관련 예제를 제시하였다.

본 연구는 트랜잭션 데이터를 분석하는데 주로 쓰여왔던 방법론인 연관분석이 가지고 있는 여러 문제를 확률 그래프 모형을 이용하여 해결하였다는 데 그 의의가 있다. 구체적으로, 일반화가 불가능하여 분류나 예측할 수 없다는 문제, 도메인 지식을 반영하기 어렵다는 문제, 특정 규칙을 일관되게 평가할 기준이 없다는 문제를 해결하였다.

하지만 본 연구에서 제안하는 방법은 이웃을 정의하기 위한 유사도 임계치를 객관적으로 설정하기 어렵다는 한계가 있다. 마지막으로 본 연구에서 제안하는 방법은 규칙을 평가하기에만 적합하다는 한계가 있는데, 이는 연관분석의 apriori 알고리즘과 비교하여 보면 빈발 항목 집합 생성에서 멈춘 것과 같다. 물론 관심 있는 다양한 확률 값을 계산할 수 있는 모델을 생성할 수 있으나, 그 관심 대상을 정하는 부분이 없으므로 규칙 생성에는 부적합하다.

따라서 추후 연구에서는 제안된 방법을 실제 분석에 응용하기 위해, 유사도 임계치를 객관적으로 설정하는 방법과 가능한 확률들을 계산하고 그 값을 내림차순으로 정렬하여 산출하는 등의 자동화된 방법을 개발하여 본 연구에서 제안된 방법을 보완한다. 이 방법은 관심 있는 다양한 확률들을 계산할 뿐 아니라 지지도나 신뢰도를 기준으로 상위 규칙을 정리하여 보여줄 수 있다.

참고문헌

Ahn, G.-S. and Hur, S. (2015), Prediction of New Customer's Degree of Loyalty of Internet Shopping Mall Using Continuous Conditional Random Field, *Journal of Korean Institute of Industrial Engineers*, **41**(1), 10-16.

Bron, C. and Kerbosch, J. (1973), Algorithm 457 : Finding All Cliques of an undirected Graph, *Communications of the ACM*, **16**(9), 575-577.

Chen, Y. and Welling, M. (2014), Bayesian structure learning for Markov random fields with a spike and slab prior, *arXiv preprint*, arXiv : 1408.2047.

eMarketer (2014), Worldwide E-commerces Sales to increase Nearly 20% in 2014.

Fuguang, B. (2015), A Novel Method of Interestingness Measures for Association Rules Mining Based on Profit, *Discrete Dynamics in Nature and Society*, 1-10.

Han, D.-S. (2009), Identification of Conserved Protein Domain Combination based on Association Rule. *Journal of KIISE : Computing Practices and Letters*, **15**(5), 375-379.

Informationweek (2012), Catalina Marketing Aims For the Cutting Edge of 'Big Data.'

Jordan, M. I. (1999), *Learning in Graphical Models*, MIT Press, Massachusetts, USA.

Kim, N.-K. (2008), Effect of Market Basket Size on the Accuracy of Association Rule Measures, *Asia Pacific Journal of Information Systems*, **18**(2), 95-114.

Kolaczyk, E. (2009), *Statistical Analysis of Network Data*, Springer, Boston, USA.

Park, D.-S. (2005), A Visualization on Data Mining for Association based on Web, *Journal of Korean Institute of Information Technology*, **3**(4), 1-9.

Rajtmajer, S. M. (2012), Introduction to Markov Random Fields.

Son, J.-E., Kim, S.-B., Kim, H.-J., and Cho, S.-Z. (2015), Review and Analysis of Recommender Systems, *Journal of Korean Institute of Industrial Engineers,* **41**(2), 185-208.

Turney, P. D. (2002), Thums up or thums down?: semantic orientation applied to unsupervised classification of reviews, *In Proceedings of the 40th annual meeting on association for computational linguistics*, 417-424.

Yang, S.-M. (2003), Discovery of Association Rules Based on items of Categorical Attribute and Quantitative Attribute, *Proceedings of Korean Institute of Information Technology*, 456-461.