JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 20, No. 6, pp. 1225-1233, Jun. 2019

# 부호화된 이분 네트워크 기반 온라인 허위 평가자 탐지

박 수 희[1*] · 노 은 하[2]
[1*]동덕여자대학교 컴퓨터학과
[2]성공회대학교 IT융합자율학부

# Online Fake Reviewer Detection Based on Signed Bipartite Network

**Suehee Pak[1*] · Eunha Rho[2]**
[1*]Department of Computer Science, Dongduk Women's University, Seoul 02748, Korea
[2]School of IT Convergence, Sungkonghoe University, Seoul 08359, Korea

**[요 약]**

온라인 평가 플랫폼은 사용자에게 유용한 제품 정보를 제공하므로 온라인 평판은 소비자 의사결정에 매우 중요한 요소이다. 이러한 평가 시스템의 영향력이 증가됨에 따라 허위 리뷰에 대한 탐색이 중요한 이슈로 다루어져 왔다. 이를 위한 여러 기법들이 다양하게 연구되어 왔으며 텍스트 기반, 행위 기반에 이어 망 기반 기법의 유용성이 발견되었다. 본 연구는 망 기반 기법을 이용한다. 평가자–평가–스토어의 관계를 부호화된 이분 네트워크로 표현하여 이를 기반으로 평가자와 평가 그리고 평가되는 대상의 관계를 표현하고 다양한 계산을 통해 평가자간 지표들을 도출하여 분석함으로써 허위 평가자를 탐색하는 방법을 제안한다. 본 연구에서는 실세계의 행태를 반영한 시뮬레이션을 통해 스토어, 평가자, 평가의 관계를 분석하여 모델의 효과성을 검증하였다.

**[Abstract]**

Online reputation is a very important factor in consumer decision making because the online evaluation platform provides useful product information to users. As the influence of these evaluation systems increases, searching for fake reviews has been dealt with as an important issue. Various techniques have been studied for this purpose, and text-based, behavior-based, and network-based techniques have been found useful. This study uses a network-based technique. Reviewer-Review-Store relationship is expressed by the signed bipartite network. A method of searching the fake reviewers is proposed by deriving and analyzing the indicators among the reviewers through various calculations. In this study, the effectiveness of the model is verified by analyzing the relationship between store, reviewer, and review through simulations reflecting real world behavior.

색인어 : 평가자 관계, 평가 네트워크, 평가자 진실성, 허위 평가

**Key word :** Fake review, Review network, Reviewer relation, Reviewer trustiness

**http://dx.doi.org/10.9728/dcs.2019.20.6.1225**



**Received** 21 May 2019; **Revised** 20 June 2019
**Accepted** 25 June 2019

***Corresponding Author; Suehee Pak**

**Tel:** +82-2-940-4587
**E-mail:** pak@dongduk.ac.kr

## Ⅰ. 서 론

온라인 제품 평가는 우리들이 구매, 음식점 방문, 서비스 선택 등의 의사 결정을 하는 데 아주 중요한 리소스가 되었다. 실제로 긍정적인 평가는 비즈니스 측면에서 상당한 금전적인 이득을 가져온다. 때문에 허위 평가 포스팅(posting)이 생겨나고 이는 점점 증가 추세에 있으며, 이러한 허위 평가의 탐지는 많은 이들의 관심사가 되어 왔다.

허위 리뷰를 탐지하기 위한 연구는 탐지 대상에 따라 분류할 수도 있고 탐지에 사용되는 특성에 따라 분류할 수도 있다. 본 논문은 특정 상품의 평가를 조작하는 평가자의 행위를 탐지하는 문제를 다룬다. 본 논문에서는 동일한 제품에 대한 평가자 사이의 평가 행위의 분석에 초점을 맞춘다.

허위 평가를 탐지하는 연구는 지난 10여 년간 언어 기반 기법, 행위 기반 기법, 망 기반 기법으로 분류하여 진행되어왔다. 본 연구는 망 기반 기법을 이용한다. 평가자–평가–스토어의 관계를 부호화된 이분 네트워크로 표현하여 이를 기반으로 평가자와 평가 그리고 평가되는 대상의 관계를 표현하고 이 네트워크에서 다양한 계산을 통해 평가자간 지표들을 도출하여 분석함으로써 허위 평가자를 탐색하는 방법을 제안한다.

평가는 긍정 혹은 부정으로 단순히 이루어진다고 가정하고 허위 평가자들은 대다수의 진실한 평가자들과 다른 평가를 내게 됨을 착안하여 평가 점수 차이 혹은 관계 분석을 통해 지표를 계산한다.

본 논문의 구성은 다음과 같다. 2장에서는 허위 평가 탐지에 관한 기존 연구들을 살펴본다. 3장에서는 직관적 가정과 관찰을 기반으로 신뢰성 판단을 위한 지수를 도출한다. 4장에서는 계산을 위한 알고리즘을 제시하고 실세계와 유사한 시뮬레이션을 통해 모델의 효과성을 검증한다. 5장에서는 결론 및 향후 연구 방향을 조망한다.

## Ⅱ. 관련 연구

허위 평가를 탐지하기 위한 연구는 2008년도부터 시작되었다. 그중 초반부터 주목을 받고 있는 기법들은 언어적, 행위적 특성을 사용한 감독된 학습(supervised learning) 모델이다. 이러한 기법들은 발전과 진화를 계속하며 텍스트 정보를 분석하거나 행위 패턴을 찾는 데 주로 집중해 왔다.

현재 미국에서 소비자에게 신뢰 받고 있는 Yelp는 기존 평가자들의 평가를 바탕으로 한 정보를 제공한다. 정직한 평가로 정확한 정보를 제공하기 위해 Yelp는 최근 몇 년간 기존 평가자들의 평가 중 허위 평가를 필터링(filtering)하는 데 주력해왔다[1]. 그러나 Yelp는 그 필터링 알고리즘을 공개하지 않으므로 Yelp에서 필터링 된 평가를 분석하여 Yelp의 필터링 방식을 알아내고자 하는 연구가 있었다[2]. 이 연구에서는 기존 연구 방법들이 실제 세계의 허위 평가를 어떻게 탐지하고 있는지를 연구하는 데 목적이 있다. 특히 대규모의 온라인 평가 사이트인 Yelp의 필터링 된, 그리고 필터링 되지 않은 평가를 가지고 실험한다. 이 연구에서는 개별 스패머를 탐지하는 것 외에 군집허위정보를 탐지하는 기법을 연구하였다. 어떤 제품에 대해서 처음 포스팅한 시각과 가장 최근 포스팅한 시각의 차이를 분석하여 이 차이가 작을수록 허위 평가자일 확률이 높다고 보았다.

[3]에서는 평가자, 평가, 스토어간의 관계를 파악하기 위해서 평가 그래프를 제안하였다. 이 그래프에서 서로 어떤 상호작용을 하는가를 파악하고 이를 분석하여 의심스러운 평가자 식별을 위한 반복적인 모델을 제안하였다.

텍스트, 행위 기반의 연구가 지속되다가 2013년에 비로소 평가 데이터의 연결성 구조에 관심을 갖는 망 기반 기법에 대한 연구가 시작되었다[4]. 이 연구에서는 특히 평가자–상품으로 이루어진 이분 네트워크에 평가의 긍정/부정 여부에 따라 링크의 부호를 +/−로 결정한 부호화된 평가 이분 네트워크(Signed bipartite review network)를 이용하여 평가자, 평가, 제품의 진실성, 정직성, 신뢰성을 계산하였다. 친구 사이가 급격하게 형성되는 양상을 평가자-평가자 망에 MRF(Markov Random Field)를 적용하여 허위 정보 탐지 프레임워크를 개발하였으며[5] 평가자-평가자 공모 관계와 평가자-속성 사이의 관계를 MRF로 모델링하였다[6]. 평가자-주소-평가 정보를 이용하여 동일한 IP 주소에서 발생하는 평가 및 평가자를 분석하는 방법도 제안되었다[7]. 그 후 그래프를 이용한 망 기반 기법의 발전적인 연구가 지속되어 왔다[8].

[9]의 연구에서는 평가자–상품평가 사이의 이분법 망에서 평가자 및 상품별 기본적인 망 특성을 분석하고 엔트로피를 분석하고 프로젝션 망에서 공모자들을 군집화(clustering)하여 리뷰 수 및 리뷰 시간 차이를 바탕으로 링크 값을 정할 경우의 군집의 특성을 분석하였다.

본 연구는 [4]의 연구에서 사용한 부호화된 평가 이분 네트워크를 그 출발점으로 하여 본 연구의 선행 연구[10]에서 사용한 모델을 정련하였다 즉, 모델의 단순성을 위하여 제한했던 선행 연구의 가정들을 현실화하고 보다 정교하고 효율적인 계산을 수행한다. 즉 평가자–스토어 간의 관계를 이분 그래프로 표현한 후 각 연결선에 긍정평가 부정평가 여부에 따라 1 혹은 −1의 값을 가중치로 두고 이를 부호화된 이분 네트워크(Signed Bipartite Network)를 사용하며, 다양한 계산을 수행하여 신뢰성을 추정하기 위한 지표를 얻어낸다.

## Ⅲ. 허위 평가 탐지 모델

본 연구에서는 평가자의 진실성(trustiness)은 진실(trusty)과 허위(fraud)로, 평가의 정직성(honesty)은 정직(honest)과 허위(fake)로, 그리고 스토어의 신뢰성(reliability)은 신뢰 있는(good)과 신뢰 없는(bad)으로 표현하기로 한다.

### 3-1 직관적 가정과 관찰

제안하는 모델을 검증하고 실험하기 위해 직관적인 가정들이 사용될 수 있는데, 본 논문에서는 선행 연구[10]에서보다 더욱 현실적인 가정들을 추가하였다. 이로써 실세계에서의 평가자들의 취향이나 행태를 보다 실질적으로 반영하였다.

- 평가는 긍정 혹은 부정으로 단순하게 이루어진다.
- 신뢰 있는 스토어는 허위 평가자를 두지 않는다.
- 신뢰 없는 스토어는 1명 이상의 허위 평가자를 둔다. 각 허위 평가자는 단 1개의 신뢰 없는 스토어와 공모한다.
- 진실한 평가자 수는 허위 평가자 수보다 상대적으로 많다.
- 신뢰 있는 스토어는 맛있을 수도 있고 맛없을 수도 있다. 신뢰 있는 스토어의 맛있거나 맛없는 비율은 다양하게 설정 가능하다.
- 허위 스토어는 맛이 없다.
- 진실한 평가자는 맛있는 스토어에는 일정 확률로 긍정 평가 점수를 주고, 맛없는 스토어에는 일정 확률로 부정 평가 점수를 준다. 일정 확률은 다양하게 설정 가능하다.
- 허위 평가자는 자신과 공모한 신뢰 없는 스토어에는 긍정 평가 점수를 주고, 그 외의 모든 스토어에는 부정 평가 점수를 준다.
- 각 평가자는 자신이 평가한 스토어에 대해 단 하나의 평가 점수를 지닌다. 한 평가자가 한 스토어를 여러 번 평가한 경우 여러 평가 점수의 평균을 최종 평가 점수로 한다.
- 참여하는 평가자는 최소 한 개의 스토어를 평가하며, 한 평가자가 평가할 수 있는 최대 스토어의 개수를 다양하게 설정할 수 있다.

위의 가정은 아직도 현실적이지 않은 부분이 있다. 평가가 긍정 혹은 부정으로 단순하게 이루어진다는 가정, 허위 스토어의 음식이 다 맛없다는 가정, 허위 평가자들은 자신이 공모하지 않은 음식점에 모두 부정 평가를 한다는 가정 등은 향후 보완할 필요가 있다.

그림 1은 앞에서 제시한 가정에 근거하여 평가자–평가–스토어 관계를 부호화된 이분 네트워크를 이용하여 표현하는 예이다. 평가는 긍정(1) 또는 부정(−1)으로 표현된다. $U_i$에서 $S_i$로의 연결 간선에서 긍정평가(1)는 실선으로 표시하고, 부정평가(−1)는 점선으로 표시하였다.

스토어 3개 중 $S_1$, $S_2$는 신뢰 있는 스토어이고 $S_3$는 신뢰 없는 스토어이며, 평가자 4명 중 $U_1$, $U_2$, $U_3$는 진실한 평가자이고 $U_4$는 $S_3$와 공모한 허위 평가자이다. $S_1$은 신뢰 있으나 맛없는 스토어, $S_2$는 신뢰 있고 맛있는 스토어이며, 신뢰 없는 스토어 $S_3$는 앞의 가정에 따라 맛없는 스토어이다. 평가는 앞에서 제시한 가정대로 이루어진다. 이 중 $U_3$의 맛없는 스토어 $S_1$에 대한 평가는 취향의 다름을 반영하여 긍정 평가가 이루어졌다고 보았다.

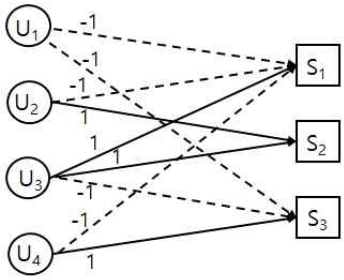


**그림 1.** 평가자(U), 스토어(S), 평가(1/−1)의 부호화된 이분 네트워크

**Fig. 1.** Signed Bipartite Network for reviews(1/-1) of Reviewer(U) and Store(S)

### 3-2 평가자 진실성지수

선행 연구에서와 같이 평가자의 진실성을 도출하였으나 본 연구에서는 다른 계산 방법을 사용한다. 선행 연구에서는 다음과 같은 추정에 의거하여 진실성지수를 계산하였다.

"진실한 평가자가 허위 평가자보다 많은 상황에서는 하나의 스토어에 대한 긍정 평가자과 부정 평가자 중 다수인 쪽이 진실일 확률이 높다."

의견이 같은 대다수에 진실성을 몰아주는 단순한 방법을 사용한 선행 연구보다 정교한 방법을 사용하기 위해 다음과 같은 직관적 추정을 이용한다.

"하나의 스토어에 대해서 다른 평가자들과의 평가 점수 차이가 클수록 평가가 허위일 가능성이 높다."

이 추정을 기반으로 다음과 같은 방식으로 평가자 간의 평가 차이를 분석하여 신뢰성을 보여주는 지수를 도출한다.

하나의 스토어에 대해서 다른 평가자들과 평가 점수 차이의 평균을 자신의 차이점수(DS: Difference Score)라 하고 모든 스토어에 대해 얻은 차이점수의 평균 점수가 그 평가자의 차이지수(DF: Difference Factor)가 된다. 이렇게 계산된 DF 값이 작을수록 평가자가 진실하다고 추정할 수 있다. 마지막 단계로, 차이지수를 아래와 같이 0과 1 사이로 정규화하여 진실성지수(TF: Trustiness Factor)를 계산한다.

$$TF = (2 - DF) / 2$$

이러한 진실성지수가 높을수록 평가자의 신뢰성이 높다고 추정할 수 있다. 표 1은 그림 1의 이분 네트워크로부터 진실성지수를 구하는 과정을 보여준다.

**표 1.** 평가자들의 차이점수, 차이지수와 진실성지수
**Table 1.** Difference Score, Difference Factor and Trustiness Factor of each reviewer

| Reviewer | Difference score | | | Difference Factor | Trustiness Factor |
|---|---|---|---|---|---|
| | $S_1$ | $S_2$ | $S_3$ | | |
| $U_1$ | 2/3 | | 2/2 | (2/3 + 2/2) /2 = 0.83 | 0.58 |
| $U_2$ | 2/3 | 0/1 | | (2/3 + 0/1)/2 = 0.33 | 0.83 |
| $U_3$ | 6/3 | 0/1 | 2/2 | (6/3 + 0/1 + 2/2)/3 = 1 | 0.5 |
| $U_4$ | 2/3 | | 4/2 | (2/3 + 4/2)/2 = 1.33 | 0.33 |

**표 2.** 평가자들의 관계점수와 관계지수
**Table 2.** Relation Score and Relation Factor for each reviewer

| Reviewer | Relation Score | | | Relation Factor |
|---|---|---|---|---|
| | $S_1$ | $S_2$ | $S_3$ | |
| $U_1$ | 3/4 | | 2/3 | (3/4 + 2/3) /2 = 0.71 |
| $U_2$ | 3/4 | 2/2 | | (3/4 + 2/2)/2 = 0.87 |
| $U_3$ | 1/4 | 2/2 | 2/3 | (1/4 + 2/2 + 2/3)/3 = 0.64 |
| $U_4$ | 3/4 | | 1/3 | (3/4 + 1/3)/2 = 0.54 |

본 연구에서는 평가를 긍정(1) 혹은 부정(-1)으로 단순화하였으나 현실세계의 평가는 일정 범위의 점수를 선택하는 방식을 사용한다. 예를 들어 5점 척도를 사용하여 1(매우 나쁨)부터 5(매우 좋음)까지의 정수 값으로 평가가 이루어진다. 이러한 경우 차이지수를 이용하여 얻은 진실성지수는 더욱 정교한 계산을 가능하게 하여 더욱 유의미한 결과를 낼 것이다.

### 3-3 평가자 관계지수

관계지수를 측정함에 있어 선행 연구[10]와 개념적으로는 유사하나 모든 평가자가 모든 스토어를 평가한다는 가정을 본 연구에서는 제외하였으므로 관계지수 계산에 있어 다른 방법이 필요하다. 또한 2차원 배열의 계산을 반복하는 선행 연구의 계산 방법을 효율적으로 개선하였다.

평가자–평가–스토어로 이루어진 이분 네트워크에서 스토어 기반의 프로젝션 네트워크(Projection Network)를 도출하였다. 그림 2는 그림 1의 이분 네트워크로부터 각 스토어에 대한 평가자–평가자 관계를 보여주는 프로젝션 네트워크를 도출한 것이다. 각 스토어에 대한 동일한 평가를 내린 평가자들 사이에 연결 간선을 삽입하고, 연결된 평가자들을 긍정평가군, 부정평가군으로 군집화하여 컴포넌트로 표현하였다. 이때 평가자들 사이의 간선과 컴포넌트 묶음을 나타내는 선은 긍정인 경우 실선으로, 부정인 경우 점선으로 표현하였다.

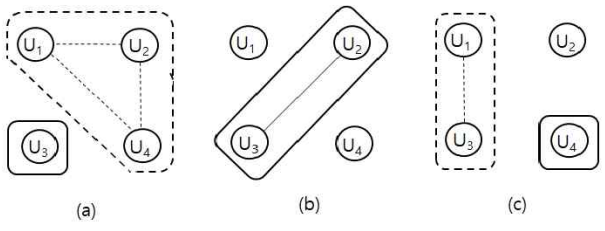


**그림 2.** 각 스토어에 대한 프로젝션 네트워크 (a)$S_1$ (b)$S_2$ (c)$S_3$
**Fig. 2.** Projection Network for each store (a)$S_1$ (b)$S_2$ (c)$S_3$

그림 2의 (a)는 스토어 $S_1$에 대해서 부정 평가를 한 $U_1$, $U_2$, $U_4$가 부정평가군을 이루고 긍정 평가를 한 $U_3$는 단독으로 긍정평가군을 이루는 것을 보여준다. (b)는 스토어 $S_2$에 대해서 $U_2$와 $U_3$가 긍정평가군을 이루고 $U_1$, $U_4$는 미평가임을 보여준다. (c)는 스토어 $S_3$에 대해서 $U_4$는 긍정평가군, $U_1$과 $U_3$는 부정평가군을 이루며, $U_2$는 미평가임을 보여준다. 평가자들간의 관계 분석을 위해 다음과 같은 직관적 추정이 가능하다.

"같은 군에 속한 평가자가 많을수록 평가자들 간의 관계성이 높다고 볼 수 있고 그 관계성이 신뢰를 대변한다."

이 관계성을 계산하기 위해서 하나의 스토어에 대해서 각 평가자는 같은 군에 속한 평가자를 전체 평가 인원수로 나눈 값을 관계점수(RS: Relation Score)로 얻게 되고 각 평가자는 모든 스토어에 대해 얻은 관계점수를 누적하여 그 평균을 낸 값인 관계지수(RF: Relation Factor)를 계산해 낼 수 있다. 이러한 관계지수가 높을수록 평가자의 신뢰성이 높다고 추정할 수 있다. 표 2는 그림 2의 프로젝션 네트워크로부터 관계지수를 구하는 과정을 보여준다.

## Ⅳ. 알고리즘 및 실험

실세계의 데이터를 대상으로 하는 연구의 선행 작업으로 본 연구에서는 시뮬레이션을 통해 위에서 제시한 모델의 타당성을 검증한다. 시뮬레이션을 위한 알고리즘에서 이용하는 값들을 아래와 같이 정의한다.

$S_i$: i번째 스토어(총 n개: $S_1$, $S_2$, $S_3$, ..., $S_n$, n = gsNb + bsNb)
$U_i$: i번째 평가자(총 m명: $U_1$, $U_2$, $U_3$, ..., $U_m$, m = guNb + totalBuNb)
gsNb: 신뢰 있는 스토어 개수(Good Store Number)
bsNb: 신뢰 없는 스토어 개수(Bad Store Number)
$GS_i$: i번째 신뢰 있는 스토어(총 gsNb개: $GS_1$, $GS_2$, ..., $GS_{gsNb}$)

$BS_i$: i번째 신뢰 없는 스토어(총 bsNb개: $BS_1$, $BS_2$, ..., $BS_{bsNb}$)
guNb: 진실한 평가자 수(Good User Number)
$buNb_i$: 신뢰 없는 스토어 $S_i$와 공모한 허위 평가자 수(Bad User Number for $BS_i$)
totalBuNb: 총 허위 평가자 수. 즉, $\sum_{i=1}^{bsNb} buNb_i$
$GU_i$: i번째 진실한 평가자(총 guNb명: $GU_1$, $GU_2$, ..., $GU_{guNb}$)
$BU_{ij}$: $BS_i$에 대한 j번째 허위 평가자(총 totalBuNb명: $BU_{11}$, $BU_{12}$, ..., $BU_{1buNb1}$, $BU_{21}$, $BU_{22}$, ..., $BU_{2buNb2}$, ..., $BU_{bsNb1}$, $BU_{bsNb2}$, ... )
$PUS_s$: 스토어 s에 대한 긍정 평가자 집합
$NUS_s$: 스토어 s에 대한 부정 평가자 집합
$ReviewNum_u$: 평가자 u의 스토어 평가 개수
$ReviewedNum_s$: 스토어 s의 평가자 수
$DS_{us}$: 평가자 u의 스토어 s에 대해 다른 평가자들 평가와의 차이 평균 점수
$DS_u$: 평가자 u의 다른 평가자들 평가와의 차이 평균 점수 합
$DF_u$: 평가자 u의 차이지수
$TF_u$: 평가자 u의 진실성지수 ($0 \le TF_u \le 1$)
$RS_{us}$: 평가자 u의 스토어 s에 대한 관계점수
$RST_u$: 모든 스토어에 대한 평가자 u의 관계점수의 합
$RF_u$: 평가자 u의 관계지수 ($0 \le RF_u \le 1$)

### 4-1 이분 네트워크를 인접행렬로 표현

앞 장에서 제시한 그림 1과 같이 평가가 이루어졌다고 가정하자. 평가 결과는 그림 3과 같은 인접행렬 BRM(Binary Review Matrix)에 저장된다. 이 때 긍정 평가는 1로 표현하고 부정 평가는 −1로 표현하며, 다음과 같은 사실을 표현한다.

신뢰 있는 스토어 $GS_1$, $GS_2$ 중 $GS_1$은 맛이 없고, $GS_2$는 맛이 있다. 신뢰 없는 스토어 $BS_1$은 맛이 없다.

$GU_1$, $GU_2$, $GU_3$는 진실한 평가자이고 $BU_{11}$은 $BS_1$과 공모한 허위 평가자이다.

진실한 평가자는 진실하게 평가하나 취향 차이에 의해서 평가가 엇갈릴 수 있다($GU_3$의 $GS_1$에 대한 평가).

$BU_{11}$은 공모한 스토어 $BS_1$ 외에는 다 부정적으로 평가한다.

| | $S_1(GS_1)$ | $S_2(GS_2)$ | $S_3(BS_1)$ |
|---|---|---|---|
| $U_1(GU_1)$ | -1 | | -1 |
| $U_2(GU_2)$ | -1 | 1 | |
| $U_3(GU_3)$ | 1 | 1 | -1 |
| $U_4(BU_{11})$ | -1 | | 1 |

**그림 3.** BRM: 평가자와 스토어간의 평가 관계를 보여주는 이분 네트워크를 인접행렬로 표현
**Fig. 3.** Adjacency Matrix representing Bipartite Review Network which shows review relationships between reviewers and stores

### 4-2 평가자의 진실성지수(TF: Trustiness Factor) 계산

BRM에 그림 3과 같은 방식으로 평가 결과가 저장되었다고 가정하고 이를 이용하여 다음과 같은 알고리즘으로 각 평가자의 진실성지수 TF를 계산한다.

Input: gsNb, bsNb, guNb, totalBuNb,
a matrix of BRM, a vector of $ReviewNum_u$
Output: a vector of $TF_u$
BRM으로부터 $DS_{us}$ 를 계산
for all s ($1 \le s \le gsNb + bsNb$)
for all u ($1 \le u \le guNb + totalBuNb$)
$DS_u = DS_u + DS_{us}$
for all u ($1 \le u \le guNb + totalBuNb$)
$DF_u = DS_u / ReviewNum_u$
for all u ($1 \le u \le guNb + totalBuNb$)
$TF_u = (2 - DF_u) / 2$

이 알고리즘은 다양한 수준의 평가 점수에 대해서 적용 가능하며, 평가 점수가 1, −1로만 구성됨을 고려하는 경우 실제 계산은 보다 효율적인 계산식을 이용할 수 있다. 이 알고리즘에 따라 그림 3의 예에 대해 TF 값을 계산한 결과는 다음과 같다.

$TF_1 = 0.583333333$
$TF_2 = 0.833333333$
$TF_3 = 0.5$
$TF_4 = 0.333333333$

여기서 TF 값이 상대적으로 높은 $U_1$, $U_2$, $U_3$는 진실한 평가자이고, TF 값이 낮은 $U_4$는 허위 평가자일 가능성이 높다고 판정할 수 있다. 그림 4는 이 평가 결과에 대한 TF 값 분포를 보여준다. 파란색 점으로 나타낸 진실한 평가자의 TF 값은 상단에 분포하고, 붉은색 x 표로 나타낸 허위 평가자의 TF 값은 하단에 분포함을 알 수 있다.

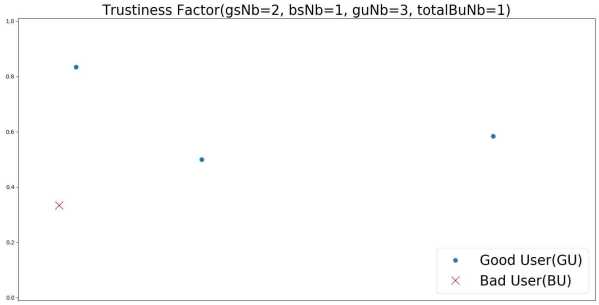


**그림 4.** 평가자 4명의 TF 분포를 나타내는 산점도
**Fig. 4.** Scatter Plot representing the distribution of TF of 4 reviewers

| | $U_1$ | $U_2$ | $U_3$ | $U_4$ |
|---|---|---|---|---|
| $U_1$ | -1 | -1 | | -1 |
| $U_2$ | -1 | -1 | | -1 |
| $U_3$ | | | 1 | |
| $U_4$ | -1 | -1 | | -1 |

(a)

| | $U_1$ | $U_2$ | $U_3$ | $U_4$ |
|---|---|---|---|---|
| $U_1$ | | | | |
| $U_2$ | | 1 | 1 | |
| $U_3$ | | 1 | 1 | |
| $U_4$ | | | | |

(b)

| | $U_1$ | $U_2$ | $U_3$ | $U_4$ |
|---|---|---|---|---|
| $U_1$ | -1 | | -1 | |
| $U_2$ | | | | |
| $U_3$ | -1 | | | |
| $U_4$ | | | | 1 |

(c)

**그림 5.** 각 스토어에 대한 프로젝션 네트워크의 인접행렬 표현 (a)$S_1$ (b)$S_2$ (c)$S_3$

**Fig. 5.** Adjacency Matrix representing Projection Network for each store (a)$S_1$ (b)$S_2$ (c)$S_3$

### 4-3 평가자-평가자 관계 분석을 위한 프로젝션 네트워크 개념 구현

스토어 s에 대한 긍정 평가자 집합 $PUS_s$와 부정 평가자 집합 $NUS_s$를 이용하여 각 스토어에 대한 평가에서의 평가자-평가자 관계를 2차원 인접행렬인 프로젝션 네트워크를 표현할 수 있다. 그림 3의 BRM을 프로젝션 네트워크로 표현하면 그림 5와 같다. 하나의 스토어에 대해서 긍정평가를 한 평가자들끼리는 연결선을 가중치로 표현한다. 긍정은 1, 부정은 −1로 가중치를 준다.

그림 5와 같은 프로젝션 네트워크를 표현한 인접행렬에서 관계지수를 계산할 수 있다. 계산의 효율성을 위해서 그림 5와 같은 인접행렬로 계산을 수행하는 대신 1차원 배열 $RST_u$를 이용하였으며 알고리즘은 다음과 같다.

```
Input:   gsNb, bsNb, guNb, totalBuNb,
         a matrix of BRM,  a vector of ReviewNum_u
Output:  a vector of RF_u
  for all s (1 ≤ s ≤ gsNb + bsNb)
    for all u (1 ≤ u ≤ guNb +  totalBuNb)
      if BRM[u][s]이 1이면   // U_u가 S_s에 긍정 평가
        u를 집합 PUS_s에 넣음
      else if BRM[u][s]이  -1이면 // U_u가 S_s에 부정 평가
        u를 집합 NUS_s에 넣음
    // RS 계산 및 누적
    for all u ∈ PUS
      RS_us = |PUS_s| / reviewedNum_s
      RST_u = RST_u + RS_us    // RS_us 누적
    for all u ∈ NUS
      RS_us = |NUS_s| / reviewedNum_s
      RST_u = RST_u + RS_us    // RS_us 누적
  for all u (1 ≤ u ≤ guNb + totalBuNb)
    RF_u = RST_u / reviewNum_u
```

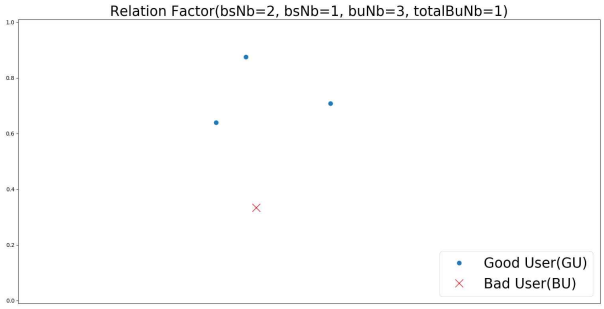


**그림 6.** 평가자 4명의 RF 분포를 나타내는 산점도

**Fig. 6.** Scatter Plot representing the distribution of RF of 4 reviewers

이 알고리즘에 따라 그림 3의 평가 예를 이용하여 각 평가자의 RF 값을 계산한 결과는 다음과 같다.

$RF_1 = 0.708333333$
$RF_2 = 0.875$
$RF_3 = 0.638888889$
$RF_4 = 0.541666667$

이 4명의 평가자 중에서, RF 값이 상대적으로 높은 $U_1$, $U_2$, $U_3$는 진실한 평가자일 가능성이 높고, RF 값이 상대적으로 낮은 $U_4$는 허위 평가자일 가능성이 높다고 판정할 수 있다. 그림 6은 이 평가 결과에 대한 RF 값 분포를 보여준다. 이 그림에서 진실한 평가자의 RF 값은 상단에 분포하고, 허위 평가자의 RF 값은 하단에 분포함을 볼 수 있다.

### 4-4 시뮬레이션

시뮬레이션을 위한 설정 및 진행은 다음과 같이 두 단계를 거친다. 첫 번째 단계에서는 본 논문에서 제시한 모델의 정당성 및 효과성을 검증하기 위한 선행 준비로서 실세계의 평가 행태를 반영하여 데이터를 발생시키는 데 집중한다. 이를 위해 3장 1절에서 제시한 관찰과 가정들을 사용하였다.

두 번째 단계에서는 앞 단계에서 발생된 평가 데이터에 본 연구의 모델을 적용한다. 이 단계에서는 앞에서 제시한 알고리즘으로 평가자의 신뢰성을 추정하기 위한 지표를 계산하고 타당성을 분석한다.

#### 1) 단계 1: 현실 세계의 평가 행태를 반영하는 데이터 발생

기본적인 입력 값은 gsNb, bsNb, guNb, $buNb_i$(for all i ∈ {1, ..., bsNb})이다. 미평가시 0, 긍정평가시 1, 부정평가시 −1의 값을 갖도록 $RATING_{ij}$(for all i ∈ {1, ..., guNb + totalBuNb}, j ∈ {1, ..., gsNb + bsNb})를 생성한다.

다음과 같은 인자들이 값을 다르게 설정함으로써 현실 세계를 반영한다.

- REVIEW_MAX: 각 평가자는 1..REVIEW_MAX개의 스토어를 평가한다.
- SAME_PREF_RATE: 진실한 평가자가 맛있는 스토어를 맛있다, 맛없는 스토어를 맛없다고 평가하는 비율이다. 진실한 평가자는 맛있거나 맛없음을 진실하게 평가하지만 맛에 대한 취향은 다를 수 있음을 감안한 인자이다.
- HOT_STORE_RATE: 신뢰 있는 스토어 중 맛있는 스토어의 비율이다.
- LOWER_TRUST_RATE: 모델을 적용한 결과 얻은 평가지수를 기준으로 하위 몇 퍼센트를 허위 평가자로 판단할 것인지를 결정하는 비율이다.

#### 1) 단계 2: 지표 계산과 분석

첫 번째 단계에서 생성한 평가 데이터 $RATING_{ij}$(for all $i \in \{1, ..., guNb + totalBuNb\}$, $j \in \{1, ..., gsNb + bsNb\}$)를 사용하여 앞에서 제시한 알고리즘에 따라 각 평가자의 TF와 RF를 구한다. 기준 비율 이하인 TF 또는 RF를 지닌 평가자를 허위 평가자로 판별한다.

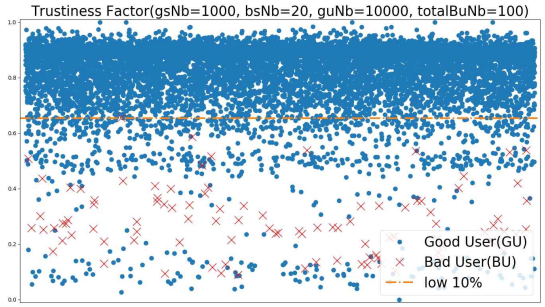


**그림 7.** 평가자 10,100명의 TF 분포를 나타내는 산점도
**Fig. 7.** Scatter Plot representing the distribution of TF of 10,100 reviewers

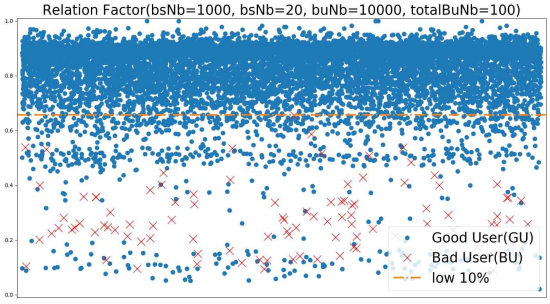


**그림 8.** 평가자 10,100명의 RF 분포를 나타내는 산점도
**Fig. 8.** Scatter Plot representing the distribution of RF of 10,100 reviewers

### 4-5 결과 분석

본 논문의 실험에서는 실제 평가 데이터를 사용하는 전 단계로 현실을 반영하도록 인위적으로 생성한 평가 데이터를 이용한다. 이러한 인공 평가 데이터에서는 진실한 평가자와 허위 평가자를 명확히 구분할 수 있다. 판별 결과를 미리 아는 상태에서 실험하여 본 논문에서 제시하는 모델이 실제로 허위 평가자를 어느 정도 판별해 낼 수 있는가를 정확히 검증할 수 있다.

다양한 크기의 평가 데이터를 생성하여 실험하여 유사한 수준으로 모델의 타당성을 확인하였다. 이 절에서 제시하는 결과 분석 예에서는 신뢰 있는 스토어 개수 1,000, 신뢰 없는 스토어 개수 20, 진실한 평가자 수 10,000, 총 허위 평가자 수 100의 크기로 평가 데이터를 생성하였다.

또한 현실 세계의 평가 행태를 다양한 측면에서 반영하여 평가 데이터를 생성했다. 이를 위해 아마존 웹 서비스(AWS)에서 제공하는 고객 평가 데이터(Amazon Customer Reviews Dataset)를 분석하여 현실 세계를 반영하도록 실험에 사용한 인자들의 값을 지정하였다. REVIEW_MAX는 30으로 설정하였다. 여기서 각 평가자의 평가 횟수는 REVIEW_MAX 이하의 랜덤 값으로 정하되, 평가 횟수가 적은 사용자가 대다수이고 평가 횟수가 많은 평가자는 소수가 되도록 가중치 랜덤(weighted random) 값으로 지정하였다. SAME_PREF_RATE는 90%로 설정하였다. 맛있는 스토어의 수가 맛없는 스토어의 수가 상대적으로 많다는 점을 반영하여 HOT_STORE_RATE는 75%로 설정하였다. 각 평가자의 허위 여부를 판별하는 기준인 LOWER_TRUST_RATE는 10%로 하여 결과를 분석하였다.

그림 7은 다음과 같은 입력 값으로 생성한 평가 데이터에 대해 TF 값을 분석한 결과를 보여준다.

gsNb = 1,000
bsNb = 20
guNb = 10,000
totalBuNb = 100

이 그림에서 0에서 1 사이의 다양한 TF 값 분포를 볼 수 있다. 파란색 점은 진실한 평가자의 TF 값이며, 붉은색 x표는 허위 평가자의 TF 값이다. 대다수의 진실한 평가자의 TF는 상대적으로 1에 가까운 쪽에 위치하고, 허위 평가자의 TF는 상대적으로 0에 가까운 쪽에 위치한다. 하위 10% 선은 이 선 아래에 위치한 평가자, 즉 TF 값이 하위 LOWER_TRUST_RATE 비율(10%)에 속하는 평가자를 허위 평가자로 판단하여 유효 평가 데이터에서 제외하기 위한 것이다. 이 실험 예에서는 100명의 허위 평가자 중에서는 1명 이 선의 윗부분에 놓여 제외되지 않고, 10,000명의 진실한 평가자 중에서는 911명이 이 선의 아랫부분에 놓여 제외되는 결과를 얻는다. 즉, TF 값이 하위 10%인 평가자의 평가를 모두 제외하는 경우, 100명 중, 99명의 허위 평가자를 걸러내므로, 효과적으로 허위 평가자를 판별할 수 있음을 보여준다.

그림 8은 그림 7과 동일한 입력 평가 데이터에 대해 RF 값을 분석한 결과이다. 대다수의 진실한 평가자의 RF는 1에 가까운 쪽에 위치하고, 허위 평가자의 RF는 상대적으로 0에 가까운 쪽에 위치한다. 하위 10% 선을 기준으로 볼 때, 100명의 허위 평가자 중에서는 0명이 윗부분에 놓이고, 10,000명의 진실한 평가자 중에서는 910명이 이 선의 아랫부분에 놓인다. 즉, RF 값이 하위 10%인 평가자의 평가를 모두 제외하는 경우, 100명 중, 100명의 허위 평가자를 걸러내므로, RF를 이용한 방법 역시 허위 평가자를 판별하는 기능에서 높은 정확도를 보여 준다.

## Ⅴ. 결 론

어떤 서비스 및 제품에 대해서 허위로 평가하여 의사 결정에 악영향을 미치는 사회적 문제에 대응하기 위한 허위 평가 탐색 기법에 대해 지난 10여 년간 많은 연구가 이루어져왔다. 본 연구는 망 기반 연구 기법을 기반으로 한다. 평가자–평가–스토어의 관계를 부호화된 이분 네트워크를 이용하여 표현한 후 이를 이용하여 다양한 계산을 시도하여 각 평가자들에 대해서 신뢰도를 분석하여 허위 평가자를 색출한다. 동일 스토어에 대한 평가의 차이를 기반으로 하여 진실성지수를 도출하였고, 동일 스토어에 대한 같은 평가를 갖는 평가자들을 묶는 방법으로 관계지수를 도출하였다. 이 두 가지 지수는 모두 평가자의 신뢰성에 대한 중요한 지표를 제공한다.

본 연구는 망 기반 허위 평가 탐지를 위한 연구로서 실제의 평가 상황을 반영한 평가 데이터 생성하고 이 데이터를 이용한 시뮬레이션을 통해서 모델의 효과성을 검증하였다. 본 연구에서는 평가자의 평가를 긍정 혹은 부정 평가로 단순화하였으나 향후 연구에서는 5점 척도의 평가를 다루는 등, 실제의 평가 상황에 더욱 근접하는 시뮬레이션을 수행하고 실제 평가 데이터에 적용하여 이 모델을 정련, 발전시킬 수 있을 것이다.

## 참고문헌


[1] A. Mukherjee, V. Venkataraman, B. Liu, and N. S. Glance (), “What Yelp fake review filter might be doing?, ” *ICWSM*, pp. 409−418, 2013.

[2] A. Mukherjee, B. Liu, and N. S. Glance, “Spotting fake reviewer groups in consumer reviews,” *IWWW*, pp. 191−200, 2012.

[3] G. Wang, S. Xie, B. Liu, and P. S. Yu, “Review graph based online store review spammer detection,” *ICDM*, pp. 1242−1247, 2011

[4] L. Akoglu, R. Chandy, and C. Faloutsos, “Opinion fraud detection in online reviews by network effects,” *ICWSM*, pp. 2−11, 2013.

[5] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, “Exploiting burstiness in reviews for review spammer detection,” *ICWSM*, pp 175−184, 2013.

[6] C. Xu,, J. Zhang, K. Chang, and C. Long, “Uncovering collusive spammers in Chinese review websites,” *CIKM*, pp. 979-988, 2013.

[7] H. Li, Z. Chen, B. Liu, X. Wei, and J. Shao, “Spotting fake reviews via collective PU learning,” *ICDM*, 2014.

[8] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Gallagher, and T. Eliassi-Rad, “Collective classification in network data,” *AI Magazine*, Vol. 29, No. 3, pp. 93-106, 2008.

[9] Suehee Pak and Eunyoung Lee, “Analysis of Crowdturfing Spam Characteristics on Online Social Media Systems,” *Journal of Digital Contents Society,* Vol. 19, No. 11, pp 2077-2084, Nov. 2018.

[10] Suehee Pak, “Signed Bipartite Network based Online Store Review Detection,” *Journal of Dongduk Information Science*, Vol. 22, 2018.

**박수희**(Suehee Pak)

1985년～1989년: 서울대학교 계산통계학과 학사
1989년～1991년: 미국 University of California, San Diego, Dept. of Computer Science(공학 석사)
1991년～1994년: 미국 University of California, San Diego, Dept. of Computer Science(공학 박사)
1995년～현재: 동덕여자대학교 컴퓨터학과 교수
※관심분야 : 소프트웨어 공학, 디지털 콘텐츠

**노은하**(Eunha Rho)

1985년～1989년: 서울대학교 계산통계학과 학사
1989년～1991년: 서울대학교 전산과학과 석사
1991년～1999년: 서울대학교 전산과학과 박사
1999년～현재: 성공회대학교 IT융합자율학부 교수
※관심분야 : 소프트웨어 공학, 디지털 콘텐츠