JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 19, No. 11, pp. 2077-2084, Nov. 2018

온라인 사회망 시스템의 집단 스팸 리뷰의 특성 분석

박 수 희 · 이 은 영
동덕여자대학교 컴퓨터학과

Analysis of Crowdturfing Spam Characteristics on Online Social Media Systems

Suehee Pak · Eunyoung Lee
Department of Computer Science, Dongduk Women's University, Seoul, Korea

[요 약]

온라인 소셜 미디어 시스템 및 다양한 리뷰 플랫폼은 사용자에게 유용한 제품 정보를 제공하여 만족스러운 제품 구매를 유도할 수 있다. 더 많은 고객이 온라인 제품 리뷰 및 등급에 의존하고 온라인 평판이 구매 결정을 내리는 중요한 요소로 자리 잡고 있다. 그러나, 리뷰 시스템의 영향력이 증가함에 따라 제품의 리뷰를 일부러 높이거나 낮추는 허위 리뷰를 게재하여 경제적 이득을 취하려는 봇 또는 스패머도 증가하고 있다. 정보가 위조되면 사용자에게 불만족스러운 경험을 초래하며, 나아가 인터넷 리뷰 시스템의 신뢰성이 손상되는 손해를 발생시킬 수 있다. 많은 연구자들이 스팸 리뷰 문제를 해결하고자 시도하였으며, 문맥, 행동 및 구조 정보를 바탕으로 스팸을 탐지하는 다양한 기법을 발견하였다. 본 논문에서는 대규모의 실제 데이터를 사용하여 군집 스팸 공격의 네트워크 특성을 분석하였다. 분석 결과 사용자-리뷰 이분법 네트워크 및 프로젝션 네트워크에서 몇 가지 고유한 특징을 발견하였다. 본 논문에서 발견한 결과는 네트워크 특징에 기반한 군집 스팸 탐지 기법을 개발하는 데 활용될 것으로 기대된다.

[Abstract]

Online social media systems and various review platforms provide useful product information to customers in purchasing most satisfactory products. As a greater number of customers relied on online product reviews and ratings, online reputation became an important factor in making purchasing decisions. However, internet review systems attracted bots and fraudsters who post promoting or demoting false reviews to achieve financial gains. Falsified information leads to unsatisfactory experiences to users and damages the credibility of internet review systems. Many researchers tackled the false review problems and discovered various spam features that can be classified into context, behavior and structure features. In this paper, we analyze the network characteristics of crowdturfing spam attacks with a large real-world dataset. Our analysis reveals several distinguishing features obtained from the user-rating bipartite network and projection networks. The results are expected to lit a light in developing detection schemes based on network features.

색인어 : 스팸 리뷰, 리뷰 시스템, 허위 리뷰, 군집 스팸 탐지, 스팸 네트워크 특성
Key word: Spam review, Review system, False review, Crowdturfing spam detection, Spam network characteristics

http://dx.doi.org/10.9728/dcs.2018.19.11.2077

Received 09 October 2018; **Revised** 22 October 2018
Accepted 20 November 2018

***Corresponding Author; Suehee Pak**

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

Ⅰ. 서 론

인터넷의 규모와 중요성이 지속적으로 증가함에 따라 Facebook등과 같은 블로그의 상품평가 및 Yelp, Amazon 등의 온라인 리뷰 시스템의 영향도 지속적으로 증가하고 있다. 예를 들어 2014년 Yelp에는 1800 만 건이 넘는 리뷰가 작성되었으며 TripAdvisor[5]는 2억 건 이상의 제품 및 서비스 평가를 보유하고 있다. 많은 소비자가 상품을 구매함에 있어 온라인 상품평을 미리 살펴보고 결정을 내리고 있으며 [1,2,3,4], 이런 점들로 인해 상품평은 광범위한 산업 분야 특히 전자 상거래 분야에서 큰 영향을 줄 수 있다. 온라인 리뷰는 세계적으로 가장 많이 사용되는 전자 상거래 시스템의 하나인 Amazon에서와 같이 전자 상거래 시스템에서 운영할 수도 있고 Yelp이나 TripAdvisor와 같이 독립적인 시스템에서 운영될 수도 있다. 특히 전자 상거래 분야에서는 고객만족도를 높이고 실소비자가 느끼는 생생한 경험과 평가를 공유함으로써 비즈니스를 개선하고 향상시키기 위해 온라인 리뷰 서비스를 자체적으로 제공하여 상품평과 구매를 직접적으로 연결하고 있다.

온라인 제품평가는 소비자들에게 생생한 정보 및 피드백을 제공한다는 순기능도 있지만 왜곡된 정보를 노출하기도 한다. 최근 연구결과에 의하면 대부분의 소비자는 물품이나 서비스를 구매함에 앞서 제품평가를 참고한다는 것이 밝혀졌고 좋은 평가가 제품의 판매량을 증가시킨다는 분석이 보고되었다. 이런 경우에 있어 허위 평가는 종종 경제적인 이득을 획득하려는 동기로부터 기원하는데 특정 제품에 관한 허위 평가를 해주면 금전을 지불하겠다고 유혹하는 광고가 나올 정도로 만연되어 있다 (그림 1)[5]. 제품 및 서비스에 과한 허위정보는 결과적으로 그런 정보를 접하고 구매결정을 하는 소비자에게 금전적으로 손해를 미치거나 만족도를 저하시킨다. 뿐만 아니라 허위정보는 궁극적으로 온라인 상품평가 시스템의 신뢰도를 낮추어 리뷰 시스템의 기능을 방해하는 결과를 초래할 수 있다. 허위 리뷰의 폐해가 심각함에 따라 리뷰 시스템의 문제점은 연구자 및 온라인 리뷰 시스템 운영자뿐만 아니라 정부기관 및 주요 언론의 주목을 받기도 하였다. 예를 들어, BBC와 New York Times는 "가짜 리뷰가 웹에서 일반적인 문제가 되고 있다“ 고 보도하였다[6]. 캐나다 정부는 온라인 리뷰 중 1/3이 허위라고 추정하였고 소비자들에게 "일반인이 올린 것으로 보이는 가짜 온라인 평가에 주의” 하라고 권유하였다[7].

허위 상품평가 및 그의 위해성이 심각해짐에 따라 이를 방어하려는 노력도 증가하고 있다. 미국 FTC는 비즈니스 운영자들이 소비자들에게 인센티브를 제공하여 리뷰를 남기도록 하는 행위를 방지하고 있다[8]. 허위 평가가 비즈니스에 심각한 위협이 될 수 있음을 인지한 Yelp, TripAdvisor 등과 같은 온라인 리뷰 시스템은 사용자들에게 리뷰 쓰기 가이드라인을 정해 홍보하고 있고 허위 리뷰를 탐지하는 비공개 자체


WANT TO EARN SOME EXTRA CASH?

ABC Restaurant would like to invite you to participate in posting an online review.

We strive to earn a **5-STAR RATING and POSITIVE REVIEWS from you!**

In return, you will receive a **$10 gift card** to a national retailer.

If you wish to participate, go to 'yelp.com/writeareview/biz/abc' to begin posting a review, and contact us at abc@abc.com.

(Please don't mention the monetary incentive on your review)

Don't miss this chance!

ABC Restaurant

그림 1. 평점 5점 리뷰를 하면 현금을 제공하겠다는 광고 (익명화됨)
Fig. 1. Ad offering cash for the 5 star rating reviews

알고리즘을 제작하여 운영하고 있다. 연구계에서도 허위 평가를 탐지하는 문제는 중요한 연구과제로 부상하고 있으며 수많은 연구결과가 발표되었다.

허위 리뷰를 탐지하기 위한 연구는 탐지 대상에 따라 분류될 수도 있고 탐지에 사용하는 특성(feature)에 의해 분류될 수도 있다. 탐지 대상은 허위 리뷰 그 자체, 허위 리뷰의 작성자, 봇 또는 시빌(sybil) 이라 불리는 허위계정, 다수의 공모자로 구성된 군집허위정보(crowdturfing: 군중의 협력 작업을 뜻하는 crowdsourcing과 대규모 허위정보 공세를 의미하는 astroturfing의 합성어)를 들 수 있다. 탐지에 사용하는 특성은 1) 평가 문장을 분석하는 언어, 2) 평가 시간 및 평점과 같은 평가자의 행위, 3) 평가자-평가자, 평가자-상품 사이의 구조적 정보를 포함한다. 이런 특성 외에 평가자의 프로필 정보나 평점 분포도 탐지 특성으로 사용하는 기법도 있다.

본 논문은 다수의 공모자가 특정 상품의 평가를 조작하는 군집허위정보 행위를 탐지하는 문제를 다룬다. 군집허위정보는 다수의 공모자가 있으므로 스패머들 사이의 관계가 존재할 것이다. 공모자들 사이의 정보를 파악하기 위해서는 공모자들 사이의 사회망 관계정보가 필요하나 많은 경우 사회망 정보 자체가 없거나 사회망 정보를 유추하기 어렵다. 따라서 본 논문에서는 사회망 정보를 추출하기 위해 평가자-상품평가 사이의 이분법 망(bipartite network)에서 평가자 사이의 관계를 추출하기 위해 망 투사(network projection) 기법을 사용하였다. 공모자들 사이의 사회적 관계 정보와 더불어 군집허위정보는 허위 평가가 시간적으로 집중되어 발생한다는 특징을 가질 것으로 판단하여 평가정보를 올리는 시간 차이를 기반으로 정보 엔트로피를 계산하였다.

본 논문은 실세계에서 얻은 대규모 평가 데이터를 바탕으로 분석을 시행하였다. 데이터는 중국 아마존에서 2010년 1월부터 2012년 8월 사이에 발생한 평가정보로 26만 명의 평가자, 6만7천여 개의 상품 그리고 백만 건이 넘는 상품 평을 포함하고 있다. 분석 결과 다음과 같은 특성이 발견하였다.

- 정상 평가자의 평가 수는 역함수를 따르고 있으나 허위 평가자의 평가 수는 역함수와 큰 차이가 있다.
- 허위 평가자의 평가는 시간적으로 집중되어 있어 정상평가자보다 더 큰 엔트로피를 가진다.

- 허위 평가자들로 구성된 클러스터는 정상 평가자와 확연하게 구별되는 클러스터 특성을 보유하고 있다.

본 논문의 구성은 다음과 같다. 2장에서는 허위 평가 탐지에 관한 기존 연구결과를 조망하였다. 3장에서는 평가자-상품평가 사이의 이분법 망에서 평가자 및 상품별 기본적인 망 특성을 분석하고 엔트로피를 분석하였다. 4장에서는 프로젝션 망(projection network)에서 공모자들을 군집화(clustering)하여 리뷰 수 및 리뷰 시간 차이를 바탕으로 링크의 값을 정할 경우의 군집의 특성을 분석하였고 5장은 향후 연구방향을 조명하고 결론을 포함하고 있다.

Ⅱ. 기존 연구 분석

허위평가 문제 및 이를 탐지하는 기법은 그 문제의 심각성으로 인해 지난 10년 동안 많은 연구가 진행된 분야이다. 기존 연구는 탐지에 사용하는 특성에 따라 1)언어 기반 기법, 2) 행위 기반 기법, 그리고 3)망 기반 기법으로 분류될 수 있다.

2-1 언어 기반 기법

언어 기반 기법은 허위평가와 진성평가에서 사용되는 언어가 다른 특성을 가진다는 사실을 기반으로 허위정보를 탐지하는 기법이다. 이 부류의 기법은 최근 발전된 word2vec[9]이나 감성분석(sentiment analysis) 연구결과를 적극적으로 활용한다. Ott et al. [10]은 허위평가의 언어적 특성뿐만 아니라 심리적인 특성을 기반으로 허위평가를 탐지하는 감독 기계학습 모델(supervised ML)을 개발하였다. 이 연구에서는 익명 평가자들이 호텔에 대한 허위평가를 올리면 돈을 지불하는 형태로 군집허위평가를 발생하였고 아마존의 'Mechanical Turk'를 사용하였다. Feng et al. [11]은 구문법 기반 탐지기법을 연구하여 자유 문법 parse 트리에서 추출한 특징이 단순한 구문론적 특징보다 탐지에 더 유용하게 사용될 수 있다는 것을 보였다. Mukherjee et al. [12]는 Yelp에서 추출한 허위평가와 우수평가를 바탕으로 언어 기반 특징과 행위기반 특징을 혼합하여 허위정보를 탐지하는 기법을 제안하였다.

2-2 행동 기반 기법

가장 빈번하게 사용되는 특성으로 평가자의 행동 패턴에서 추출한 특징을 활용한다. 대표적인 행동기반 메타 특성으로 평점 분포, 단어 사용의 특성과 같은 평가 컨텐츠, 평가 대상의 특징 (가격대, 브랜드 등) 등이 있다. Jindal and Liu [13]은 아마존의 리뷰 중 중복된 리뷰를 허위 평가로 판단한 실측자료(ground truth)를 바탕으로 36개의 행동기반 특성을 감독 학습 방법으로 추출하였다. 이 연구에서는 허위 리뷰를 세 가지 형태 (허위정보, 브랜드 이름 알리기, 무의미성)로 분류해서 각 형태의 공격에 알맞은 특성들을 추출하였다. Jindal et al.[14]은 위 연구결과를 더욱 발전시켜 평가자의 평점 분포와 평가하는 상품 브랜드의 분포를 바탕으로 허위 평가를 탐지하는 기법을 제안하였다. Mukherjee et al.[15]은 자율학습 베이지안 추론 프레임워크를 사용하여 허위 평가자를 탐지하는 기법을 개발하였다. Xie et al.[16]은 오랜 시간동안 평가자의 행동을 추적하여 평균 평점, 평가 횟수, 단일 평가 (singleton review) 횟수 통계를 바탕으로 허위 평가자를 탐지하는 기법을 제안하였다. 개별 스패머를 탐지하는 것 외에도 군집허위정보를 탐지하는 기법이 연구되었고[17, 18] 이들 기법은 공모된 군집 활동으로 발생하는 다양한 행동 패턴을 탐지 특징으로 활용한다.

2-3 망 기반 기법

Wang et al.[19]은 리뷰 시스템에서 발생하는 평가정보 및 행위를 평가자-평가-상품 이라는 망으로 표현하여 망 특성을 분석하여 평가자 및 평가의 신용도(trustiness), 평가의 정직성(honesty) 및 상품의 신뢰도(reliability) 등을 웹 정보의 중요도를 결정하는데 사용되는 HITS (Hyperlink-Induced Topic Search) 기반 모델을 사용하여 수학적으로 계산하는 방법을 제안하였다. Akoglu et al. [20]는 평가자-상품으로 이루어진 평가정보 이분법적 망을 평가가 긍정/부정 여부에 따라 링크의 부호를 +/-로 결정한 부호화된 망(signed network)으로 변환하였고 이를 바탕으로 MRF(Markov Random Field)에 기반 한 허위정보 탐지 프레임워크를 개발하였다. MRF 기법은 망 기반 탐지기법에서 많이 활용되는 기법으로 [21]에서는 급격한 친구사이 형성을 평가자-평가자 망에 MRF를 적용하여 탐지하였고 [18]에서는 평가자-평가자 공모 관계와 평가자-속성 사이의 관계를 MRF로 모델링하였다. Li et al.[22]은 평가자-주소-평가정보를 바탕으로 동일한 IP 주소에서 발생한 평가 정보 및 평가자를 탐지하는 비교적 단순하면서도 효과적인 방법을 제안하였다. Ye와 Akoglu[23]는 그래프 기반 기법을 활용하여 허위 평가자의 망 발자국(network footprint)을 모델링하는 기법을 제안하였다.

Ⅲ. 군집허위평가의 기본적 망 특성 분석

본 논문에서는 대규모 실세계 평가정보를 바탕으로 망 특성을 분석한다. 사용한 데이터 2010년 1월 1일부터 2012년 8월 29일까지 중국 아마존에서 추출한 평가 자료로서 허위평가라고 레이블된 실측자료가 존재한다. 연구 목적상 단일 평가를 받은 상품은 데이터에서 제외하였고 전처리후의 데이터 규모는 표 1과 같다. 이 장에서는 평가정보를 그림 2와 같이 평가자-상품 사이의 이분법 망으로 표현해서 평가자, 상품별 링크 수 분포(degree distribution), 평가자, 상품의 엔트로피 등과 같은 기본 특성을 분석하였다.

표 1. 연구 대상 데이터
Table 1. Research Data

total reviews	reviewers	products	false reviewers	false products
1,022,223	261,292	67,044	1,930	19,019

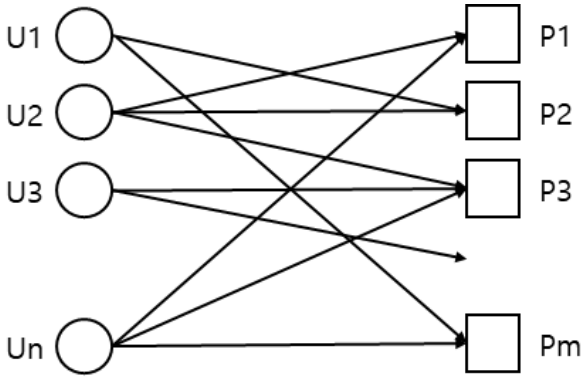


그림 2. 평가자(U), 상품 이분법 망
Fig.2 Reviewer(U), Product bipartite network

그림 2는 평가자 수, 상품 수는 각각 n, m이며 U1이라는 평가자는 P2, Pm이라는 상품에 평가를 남긴 것이고 평가자 U2는 상품 P1, P2, P3 에 평가를 남긴 것이다.

대부분의 사회 망은 역함수(power law distribution) 특성을 가진다고 보고되어 있다. 그림 3은 평가자가 남긴 평가수(그림 3 (a)) 및 상품이 받은 평가 수 (그림 3 (b))를 log-log 축척도로 표현한 것이다. 그림 3 (a)에서 정상 평가자는 역함수 분포를 따르는데 반해 허위 평가자는 역함수 분포를 따르지 않는다는 것을 알 수 있다, 즉, f(k)가 k개의 평가를 남긴 평가자의 비율이라고 한다면 정상 평가자의 분포는 역함수 파라미터 c에 의해 $f(k) \propto k^{-c}$라는 형태로 모델링될 수 있다. 기존 연구에서는 역함수 분포는 탐지 특성으로 일반적으로 사용되지 않았으므로 이 특성을 활용하여 허위 평가자를 탐지하는 새로운 기법을 도출할 가능성이 있다.

그림 3 (b)는 동일한 분석을 상품이 받은 평가 수를 기반으로 수행한 결과이다. 이 분석에서 한 번이라도 허위평가를 받은 상품은 허위상품이라고 분류하였다. 그림에서 허위상품과 정상상품은 모두 역함수 특성을 가진다는 것을 관찰 할 수 있다. 단, 정상상품의 역함수 파라미터 c가 허위상품의 파라미터보다 상당히 크다는 것을 관찰하였다. 허위상품은 허위평가를 받기도 하지만 정상적인 평가를 받기도 한다. 이런 이유로 허위상품도 역함수 분포 특성을 가질 수 있다고 가정할 수 있다. 이 가정을 확인하기 위해 허위상품이 받는 평가 중 허위평가를 제외한 정상평가 만을 가지고 분석을 실행하였다, 그림 3 (b)에서 초록색 점과 선은 허위평가를 제외한 허위상품의 평가 수를 분석한 것이다. 그림에서와 같이 허위 상품의 평가 수도 역함수 분포를 따르고 있으며 기울기가 정상상품의 평가 수 분포의 기울기와 유사해 짐을 분석하였다.

정상상품의 리뷰 수, 허위상품의 리뷰 수, 그리고 허위 리뷰를 제외한 허위상품의 리뷰 수의 역함수는 각각 2.19, 1.35, 1.59 이다. 대규모 데이터에서 많이 관찰할 수 있는 역함수의 파라미터 값이 2와 3 사이인 것으로 보고되어 있으므로 정상상품의 역함수는 통상적으로 관찰할 수 있다. 역함수의 GOF(Goodness Of Fit)를 표현하는 R^2 값은 각각 0.95, 0.93, 0.92 로 큰 차이는 없다.

평가자 및 허위상품 기반의 역함수 특성 분석과 더불어 정보 엔트로피(entropy)를 분석하였다. 엔트로피는 다양한 측면에서 측정할 수 있는데 본 논문에서는 허위 리뷰가 특정 상품에 치중되어 발생할 것이라는 가정과 허위 리뷰들이 단기간 동안 집중되어 발생할 것이라는 가정을 검정하기 위한 엔트로피를 계산하였다. 허위 리뷰가 특정 상품에 치중할 것이라는 가정을 검증하기 위하여 하루를 단위로 해서 각각의 상품이 받는 리뷰 수를 집계하여 이를 바탕으로 엔트로피를 계산하였다. 그림 4는 매일 계산한 엔트로피를 시계열로 표현한 것이다. 엔트로피는 리뷰 수의 일주일 평균을 원시 데이터로 사용하여 계산하였으며 데이터의 연속성을 유지하기 위해 인접 데이터들끼리는 2일 간 겹치도록 하였다. 그림 4에서 허위평가 수는 시간이 갈수록 증가하며 엔트로피도 이에 따라 증가한다는 것을 관찰 할 수 있다. 허위 리뷰 수와 엔트로피와의 상관관계는 시각적으로 관찰하기 어려우므로 지역화(localization) 기법 등을 적용한 분석이 요구된다.

허위 리뷰, 특히 군집허위 리뷰는 압축된 시간 내에서 발생할 것이라는 가정을 검증하기 위해 각 상품별로 리뷰의 도착시간 차이 (inter-arrival time)를 바탕으로 엔트로피를 계산하였다. 어느 상품이 k 개의 리뷰를 받는다면 (k-1) 개의 도착시간 차이가 발생하며 이를 분포화하여 엔트로피를 계산하였다. 그림 5는 허위평가를 이상 받은 763 개 모든 상품과 정상상품 중 763 개를 무작위로 샘플링하여 이들의 엔트로피의 분포를 표현한 것이다. 그림에서 허위상품이나 정상상품 모두 엔트로피는 증가하다가 감소하는 경향이 있다는 것을 관찰할 수 있다. 정상 상품은 허위 상품에 비해 엔트로피가 낮아 리뷰가 보다 골고루 분포되어 있다는 것을 확인할 수 있다.

Ⅳ. 허위 평가자 관계 분석

본 논문에서는 평가자 및 상품 각각의 특성과 더불어 평가자-평가자 사이의 관계도 분석하였다. 평가자-상품으로 이루어진 이분법 망은 허위 평가자 사이의 관계를 표현하지 못하므로 평가자들 사이의 관계를 도출하기 위한 상품 기반의 프로젝션 망(projection network)을 도출하여 평가자-평가자 사이를 분석하였다. 이분법 망에서 상품 기반 프로젝션 망을 도출하는 예는 그림 6과 같다. U1과 U3은 상품 P1을 공동으

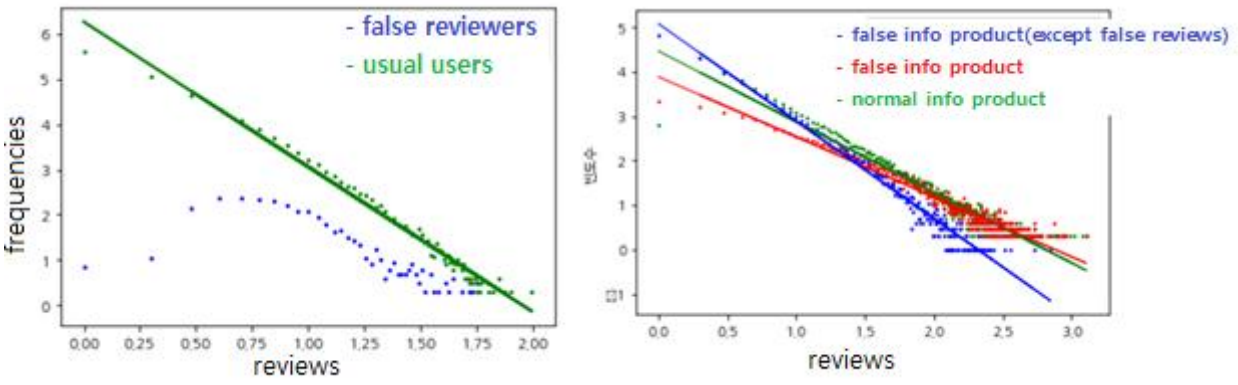


그림 3. 평가자 및 상품의 평가 수 분석
(a) (왼쪽) 평가자의 평가수
(b) (오른쪽) 상품이 평가받은 수

Fig. 3. Reviewer and product reviews analysis
(a) (left) Number of reviewer's reviews
(b) (right) Number of product reviews

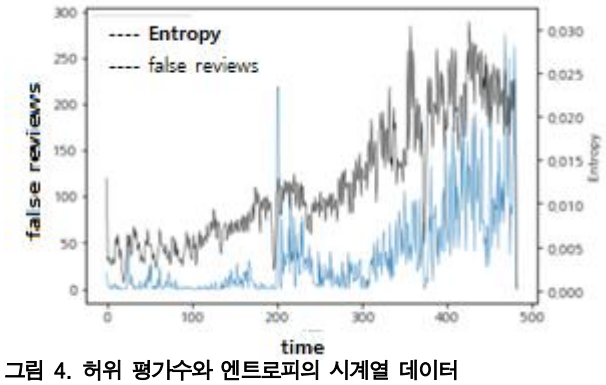


그림 4. 허위 평가수와 엔트로피의 시계열 데이터

Fig. 4. Time series data of number of false reviews and entropy

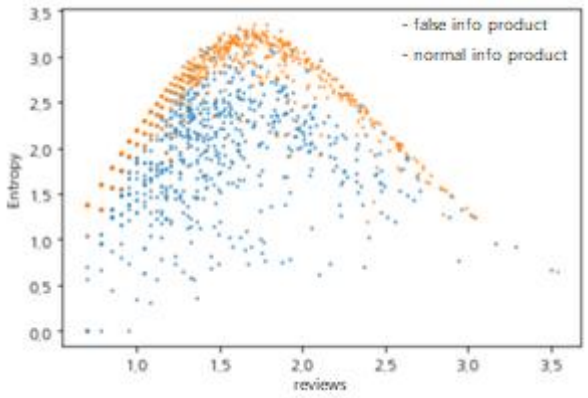


그림 5. 정상상품과 허위상품의 평가 수에 따른 엔트로피

Fig. 5. Entropy for reviews of good product and false product

로 평가하였으므로 프로젝션 망에서 링크로 연결된다. U2와 U4는 공동으로 평가한 상품이 P2, P3 두 개이므로 두 개의 링크로 연결되고 여기에서는 링크 값이 2인 링크로 표현하였다.

본 연구에서는 각 상품별로 프로젝션 망을 생성하여 허위 평가자들 사이의 특성을 분석하였다. 어느 상품 하나를 기반으로 프로젝션 망을 생성하면 모든 평가자가 서로 직접 연결되는 완전 그래프(complete graph) 형태를 가질 것이다. 본 연구에서는 허위 평가자들 사이의 특성을 분석하기 위해 링크 값을 두 가지 방법으로 부여하였다.

- 방법 1(W1): 평가자 u, v 가 평가한 상품 수를 각각 n(u), n(v) 라고 하면 링크 값은 $n(u)\cdot n(v)$ 로 결정된다. 평가자의 활동이 활발할수록 링크 값이 커진다.

- 방법 2(W2): 평가자 u, v 가 평가한 날짜를 각각 t(u), t(v) 라고 하면 링크 값은 $1/(|t(u)-t(v)|+1)$ 로 결정된다. 평가자가 평가한 시간이 가까울수록 높은 링크 값을 가진다. 이 링크 값은 시간적으로 밀집된 평가를 하는 평가자들을 판별할 수 있을 것이다.

본 연구에서는 허위평가 공모자들로 구성된 클러스터를 구성하고 클러스터 내부의 링크 값과 클러스터 외부의 링크 값을 비교하여 허위 평가자로 구성된 클러스터가 어떤 특징을 가지고 있는지 분석하였다. 보통 소셜망 분석에서 사용하는 기법은 클러스터 멤버들 사이는 더욱 밀접하게 연결되어 있고 외부와의 연결은 많지 않다는 특성을 나타내는 메트릭(metric)을 사용한다. 또는 컨덕턴스(conductance) 라는 메트릭도 클러스터의 특성을 표현하는 도구로 많이 사용되고 있다. 그러나 본 논문에서 다루는 프로젝션 망은 모든 노드가 서로 연결된 완전 그래프 형태이므로 기존 방법을 사용하는 것이 불가능하고 새로운 메트릭을 개발해야 한다.

본 연구에서는 허위 평가자의 특성을 다양한 측면에서 발견하기 위해 다음과 같은 세 가지 메트릭을 정의하였다.

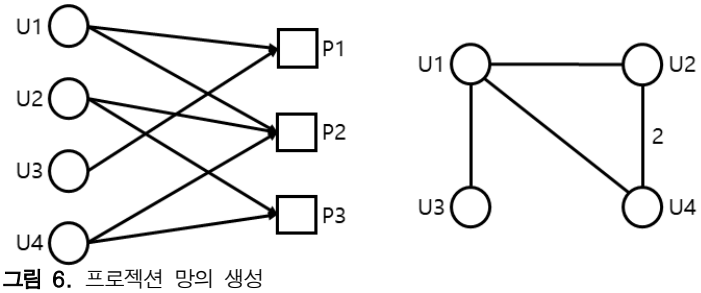


그림 6. 프로젝션 망의 생성
(a)(왼쪽)이분법 망
(b)(오른쪽)이분법 망에서 상품을 기반 생성된 프로젝션 망

Fig. 6. Generation of projection network
(a)(left) bipartite network
(b)(right) projection network
based on product of bipartite network

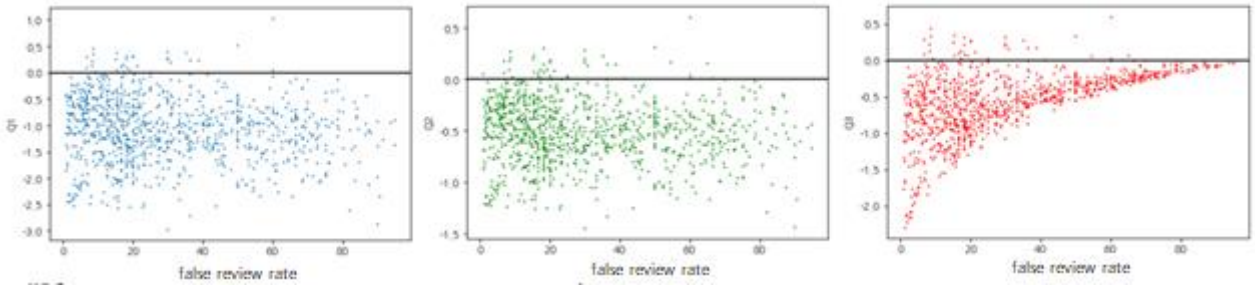


그림 7. 링크 값으로 W1을 사용한 클러스터의 특성
(a)(왼쪽) Q1
(b)(가운데) Q2
(c)(오른쪽) Q3

Fig. 7. Cluster characteristics of using W1 for link value
(a)(left) Q1
(b)(center) Q2
(c)(right) Q3

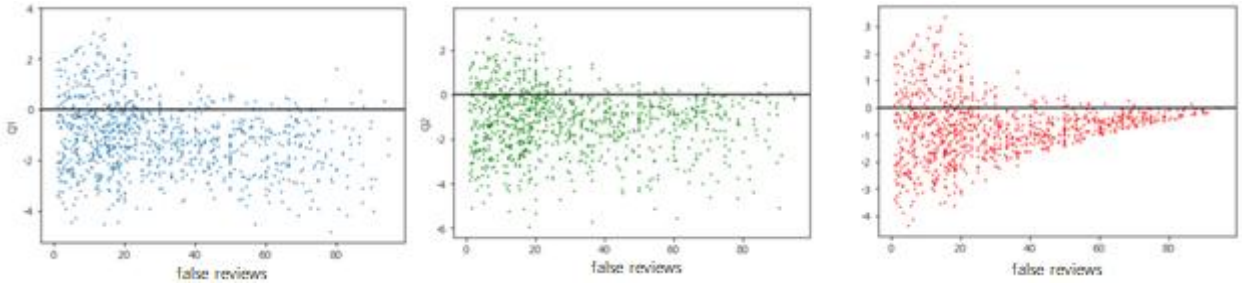


그림 8. 링크 값으로 W2을 사용한 클러스터의 특성
(a)(왼쪽) Q1
(b)(가운데) Q2
(c)(오른쪽) Q3

Fig. 8. Cluster characteristics of using W2 for link value
(a)(left) Q1
(b)(center) Q2
(c)(right) Q3

(1)Q1: (클러스터에 포함되지 않는 노드들 사이의 링크 평균값) / (클러스터 내부 링크의 링크 평균 값)
(2)Q2: (클러스터 내부 노드와 외부 노드를 연결하는 링크의 평균값) / (클러스터 내부 링크의 링크 평균 값)
(3)Q3: (전체 링크의 평균 값) / (클러스터 내부 링크의 링크 평균 값)

그림 7과 8은 W1 과 W2의 메트릭을 링크 값으로 사용했을 때의 클러스터의 특성 Q1, Q2, Q3을 계산하여 허위 평가수별로 표시한 것이다. 클러스터의 특성을 나타내는 Q 값이 널리 분포하므로 Q 값에 로그를 취하였다. 그림 7에서 대부분의 허위 상품은 0 이하의 값을 가지며 이는 대부분의 허위 상품의 Q 값이 1 이하로서 허위 평가자들이 많은 상품을 리뷰한다는 것을 보여주고 있다. 메트릭으로 W2를 사용한 그림 8도 대부분의 허위상품이 0 이하에 분포(즉, Q 값이 1 이하)하여 허위평가자들이 시간적으로 군집된 평가를 하는 것을 관찰하였다.

Ⅴ. 향후 연구 및 결론

본 논문에서는 다수의 허위 평가자가 공모하여 특정 상품의 명성을 부당하게 올리거나 내리는 군집허위평가 공격을 대상으로 일차적인 망 특성과 허위 평가자들 사이의 관계를 분석하였다. 분석에 사용된 데이터는 2010년부터 2012년 8월 사이에 중국 아마존에서 관찰한 평가 데이터로서 백만 개 이상의 평가를 포함하고 있다. 평가자가 특정 상품을 평가하면 이 둘 사이에 링크가 형성되고 모든 평가를 망으로 표현하면 이분법 망으로 표현될 수 있다. 본 연구에서는 이분법 망의 평가자 별 평가 수, 그리고 상품별 평가수를 대상으로 평가 수 분포가 역함수에 따르는지 분석하였다. 분석 결과 정상 평가자의 평가 수는 역함수를 잘 따르나 허위 평가자는 역함수와 크게 차이나는 것을 발견하였다. 또한 상품별 평가수를 분석한 결과 정상상품, 허위상품 모두 역함수를 어느 정도 근사하게 따르나 정상상품의 역함수 기울기(파라미터)가 허위상품의 기울기보다 매우 크다는 것을 발견하였다. 평가수의 역함수 뿐만 아니라 일일 평가 및 평가들 사이의 도착시간 차이를 바탕으로 정보 엔트로피를 구해 차이점을 분석하였다. 분석결과 허위상품의 도착시간 차이 엔트로피는 정상상품의 도착시간 차이 엔트로피보다 크다는 것을 발견하였고 이는 허위상품의 평가가 시간적으로 집중되어 발생한다는 것을 탐지할 수 있는 좋은 지표임을 발견하였다.

이분법 망은 평가자와 평가자 사이의 관계를 나타내기 어려우므로 상품을 중심으로 프로젝션 망을 생성하였다. 어느 한 상품을 기반으로 프로젝션 망을 만들 경우 모든 평가자가 연결되는 완전 그래프가 생성된다. 클러스터의 성질을 표현하는 기존 메트릭은 완전 그래프에서 사용할 수 없으므로 각 평가자의 평가수를 바탕으로 하는 링크 값을 결정하는 방법과 평가 사이의 차이를 바탕으로 링크 값을 결정하는 두 가지 방법을 제안하였다. 또한 클러스터의 성격을 표현하는 세 가지 방법을 제안하여 클러스터의 성격을 분석하였다. 분석결과 허위 평가자로 구성된 클러스터는 일반 평가자에 비해 더 많은 리뷰를 하며 평가가 시간적으로 집중되어 있다는 것을 발견하였다.

본 연구는 허위평가를 탐지하는 새로운 방법을 제안하기 위한 기초적 연구 성격을 가지고 있다. 본 연구결과를 바탕으로 다양한 랜덤 포레스트 등 다양한 기법을 사용하여 허위평가자 및 허위상품을 탐지하는 기법을 연구 중에 있다. 또한 본 연구팀은 평가자를 기반으로 하는 프로젝션 망을 구성하여 허위상품들 사이의 특성을 분석하는 연구도 동시에 추진하고 있다.

감사의 글

이 논문은 2017년도 동덕여자대학교 학술연구비 지원에 의하여 수행된 것으로서, 관계부처에 감사드립니다.

참고문헌

[1] P. A. Dow, L. A. Adamic, and A. Friggeri, “The Anatomy of Large Facebook Cascades,” ICWSM, pp. 145–154, 2013.

[2] Y. Matsubara, Y. Sakurai, B. A. Prakash, L. Li, and C. Faloutsos, “Rise and fall patterns of information diffusion: model and implications,” KDD, pp. 6–14, 2012.

[3] B. Hooi, N. Shah, A. Beutel, S. Gunneman, L. Akoglu, M. Kumar, D. Makhija, and C. Faloutsos, “Birdnest: Bayesian inference for ratings-fraud detection” *SDM,* Vol. 16, pp. 495-503. SIAM, 2016.

[4] Daniel Y. T. Chino, Alceu F. Costa, Agma J. M. Traina, and Christos Faloutsos, Proceedings of the 2017 SIAM International Conference on Data Mining, pp. 108-116, 2017

[5] Choi, Sungwoo, et al., "The Role of Power and Incentives in Inducing Fake Reviews in the Tourism Industry." *Journal of Travel Research*, 2016.

[6] Lau RY, Liao SY, Kwok RCW, Xu K, Xia Y, Li Y, “Text mining and probabilistic language modeling for online review spam detecting.” *ACM Trans Manage Informaion System*, Vol. 2, No. 4, pp. 1–30, 2011

[7] http://www.competitionbureau.gc.ca/eic/site/cb-bc.nsf/eng/03782.html

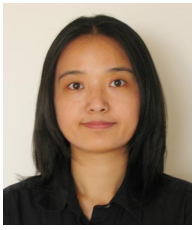
[8] http://blog.grade.us/online-review-incentives/, Online Review Incentives: Smart Marketing or Asking for Trouble? 2016.

[9] Mikolov, Tomas; et al., "Efficient Estimation of Word Representations in Vector Space", arXiv:1301.3781

[10] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, “Finding deceptive opinion spam by any stretch of the imagination,” ACL, pp. 309–319, 2011.
[11] S. Feng, R. Banerjee, and Y. Choi, “Syntactic stylometry for deception detection,” ACL, 2012.
[12] A. Mukherjee, V. Venkataraman, B. Liu, and N. S. Glance, “What Yelp fake review filter might be doing?,” ICWSM, 2013.
[13] Jindal, Nitin, and Bing Liu, "Opinion spam and analysis," Proceedings of the 2008 International Conference on Web Search and Data Mining, ACM, 2008.
[14] N. Jindal, B. Liu, and E.-P. Lim, “Finding unusual review patterns using unexpected rules,” CIKM, 2010.
[15] A. Mukherjee, A. Kumar, B. Liu, J. Wang, M. Hsu, M. Castellanos, and R. Ghosh, “Spotting opinion spammers using behavioral footprints,” KDD, 2013.
[16] S. Xie, G. Wang, S. Lin, and P. S. Yu, “Review spam detection via temporal pattern discovery,” KDD, 2012.
[17] A. Mukherjee, B. Liu, and N. S. Glance, “Spotting fake reviewer groups in consumer reviews,” IWWW, 2012.
[18] C. Xu, J. Zhang, K. Chang, and C. Long, “Uncovering collusive spammers in Chinese review websites,” CIKM, pp. 979–988, 2013.
[19] G. Wang, S. Xie, B. Liu, and P. S. Yu, “Review graph based online store review spammer detection,” ICDM, 2011.
[20] L. Akoglu, R. Chandy, and C. Faloutsos, “Opinion fraud detection in online reviews by network e↵ ects,” ICWSM, 2013.
[21] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, “Exploiting burstiness in reviews for review spammer detection,” ICWSM, 2013.
[22] H. Li, Z. Chen, B. Liu, X. Wei, and J. Shao, “Spotting fake reviews via collective PU learning,” ICDM, 2014.
[23] 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.

박수희(Suehee Pak)

1989년 : 서울대학교 계산통계학과 (학사)
1991년 : 미국 University of California, San Diego Dept. of Computer Science(공학 석사)
1994년 : 미국 University of California, San Diego Dept. of Computer Science(공학 박사)
1995년～1998년: 동덕여자대학교 컴퓨터학과 조교수
1999년～2007년: 동덕여자대학교 컴퓨터학과 부교수
2008년～현재: 동덕여자대학교 컴퓨터학과 교수
<관심분야> 소프트웨어 공학, 멀티미디어 및 디지털 콘텐츠



이은영(Eunyoung Lee)

1996년: 고려대학교 전산과학과 (학사)
1998년: 고려대학교 전산과학과 (전산학 석사)
2004년: 미국 Princeton University (전산학 박사)
2005년～2010년: 동덕여자대학교 컴퓨터학과 조교수
2011년～2016년: 동덕여자대학교 컴퓨터학과 부교수
2017년～현재: 동덕여자대학교 컴퓨터학과 교수
<관심분야> 소프트웨어 보안, 프로그래밍 언어, 클라우드 컴퓨팅