특집 03

온라인 공간에서 비정상적 정보 탐지 기술의 동향 및 발전 방향

이시형 (서울여자대학교)

목 차 »

1. 서 론

온라인 공간은 의견 교환을 통해 사용자들 간 상호작용 및 통신이 이루어지는 가상공간을 의미한다. 예를 들어, 트위터[1], 다음 아고라[2], 네이버 뉴스[3] 등은 사회, 정치 관련 의견 공유가 이루어지는 온라인 공간이며, 아마존[4] 등은 상품에 대한 의견 공유가 이루어지는 온라인 공간으로 볼 수 있다. 온라인 공간은 스마트 단말기 사용자의 증가로 정보 유통에 있어 그 영향력이 갈수록 증가하고 있으며, 이러한 공간에서의 사회적 상호작용은 이미 생활의 일부로 인식되기도 한다[1].

이와 같은 온라인 공간의 영향력을 악용해 조직적으로 비정상적 정보를 유포하는 사례 또한 증가 추세에 있다. 비정상적 정보는 사용자가 허위 정보를 사실로 인식하도록 유도하기 위해 의도적으로 유포되는 정보를 말하며, 실례로 경쟁사 제품의 판매량을 감소시키기 위해 비방사실을 유포한 '처음처럼 소주 내 알칼리 환원수 유해성 논란'[5], 정치적 목적을 위해 여론을 조작하고자 했던 '국정원 선거 개입 사건'[6], 사건 근원지에 대한 허위 사실이 유포되었던 '천안함 사건'[7] 등이 있다. 이러한 비정상적 정보는 종종 비판 없이 받아들여져, 10% 이상의 사용자에게서 실제 행동의 변화를 일으키기도 하며[2], 이후 다른 허위 사실의 유포를 위한 근거로 재사용되기도 한다[3].

온라인 공간의 정보 파급 속도 및 범위를 고려할 때 비정상적 정보의 유포를 미연에 방지하거나, 발생 시 조기 발견 및 차단해 주는 방법에 대한 연구가 시급하다고 할 수 있다. 따라서 본 논문에서는 현재까지 제안된 비정상적 정보의 탐지 기법에 대해 비교 분석하고, 앞으로의 연구 방향에 대해 알아보도록 하겠다.

1) www.twitter.com
2) agora.media.daum.net
3) news.naver.com
4) www.amazon.com
5) www.hyundaenews.com/sub_read.html?uid=3389
6) ko.wikipedia.org/wiki/대한민국_국가정보원_여론_조작_사건
7) http://ko.wikipedia.org/wiki/PCC-772_천안

2. 비정상적 정보의 유형

이번 장에서는 비정상적 정보 중 가장 많은 비중을 차지하는 두 가지 유형을 정보를 유포하는 목적에 따라 상업적인 유형(3.1절)과 정치적인 유형(3.2절)으로 나누어 소개하도록 한다.

2.1 상업적 목적의 비정상적 정보

특정 상품, 서비스, 브랜드에 대해 사용자들이 긍정적인 인식 또는 부정적인 인식을 갖도록 허위 후기 등을 게재하는 경우를 말한다. 미국의 대표적인 상점 검색 서비스인 Yelp에서는 30% 이상의 후기가 대가를 받고 인위적으로 작성되었음이 보고되었으며[4] 애플의 iPhone 앱스토어에서도 유사한 후기가 다수 발견되었다[5]. 이러한 허위 후기는 경쟁사 제품의 판매량이나 브랜드 가치를 하락시켜 금전적으로 막대한 손실을 입히기도 한다.

2.2 정치적 목적의 비정상적 정보

선거에서 유리한 위치를 선점하거나 민심을 얻기 위한 목적으로 특정 정치인, 정당에 대해 그릇되거나 과장된 사실을 유포하는 경우를 말한다. 예를 들어, 러시아에서 있었던 반정부 시위기간 동안에는 시위를 지지하는 시민이 많았음에도 불구하고, 트위터에서는 반정부 성향의 글을 거의 볼 수 없었다. 이는 압도적으로 많은 정부 지지글이 지속적으로 게재되었기 때문인데, 이를 위해 수천 개 이상의 봇과 트위터 계정이 사용되었음이 알려졌다[6]. 정치적 목적의 비정상적 정보로 인해 일부 시민의 정치적 선호도가 반전되기도 하며, 근소한 차이를 보이는 선거에서는 후보의 당락에도 영향을 미친다.

3. 비정상적 정보의 탐지 연구 동향

본 장에서는 탐지에 활용하는 비정상 정보의 특징에 따라 탐지 방법을 네 가지로 나누어 소개한다. 제시된 특징들을 이용한 탐지시스템은 게재된 의견 또는 이를 게재한 사용자에서 나타내는 특징을 입력으로 받아, 정상 또는 비정상으로 분류한다. 분류 기준은 주로 기계학습(Machine Learning) 기법을 사용하여 생성하며, SVM (Support Vector Machine) 알고리즘이 가장 많이 사용되었다[7]. 보다 정확도를 높이기 위해 Boosting[8]이나 Bagging[9]이 사용되기도 한다. Boosting은 잘못 분류된 객체에 가중치를 주고 분류 기준을 점진적으로 개선해 나가며, Bagging은 다수의 분류 기준을 만들어 평균을 취함으로써 편차가 작은 기준을 생성한다.

3.1 게재 내용, 시간의 비정상 정도 활용

비정상정보 탐지 기법에서 가장 빈번하게 사용되는 특징은 게재되는 내용에 나타나는 특징 및 게재가 일어난 시기와 관련된 특징이다.

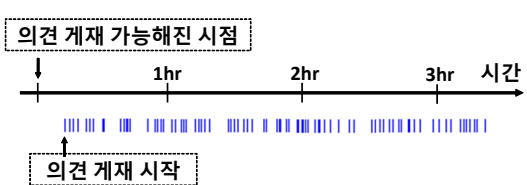
게재되는 내용을 이루는 기본 요소는 구성 단어의 집합이다. 상품에 대한 평가와 같이 호감 정도를 정해진 정수범위에서 선택할 수 있는 경우에는 평가등급(rating)도 포함된다 (e.g., 별 1~5개 범위에서 하나를 선정). 위와 같은 요소가 평균값에서 벗어나는 정도가 클수록 비정상일 가능성이 높다. 예를 들어, 상품을 고의적으로 비방하는 경우 평가등급이 평균등급보다 낮아진다[10]. 또한, 장단점을 함께 기술하기보다 무조건적으로 비판한 경우에는 부정을 나타내는 단어의 빈도가 평균 이상으로 높아진다.

유사한 내용이 반복하여 나타나는 것도 비정상

적 정보에 자주 보이는 특징이다[11,12]. 이는 보다 많은 사용자가 비정상 정보에 노출되고 결과적으로 영향 받도록 하기 위함이다. 거의 동일한 내용이 수천회 이상 대량으로 게재되는 경우도 드문 일은 아니다.

게재가 일어난 시기와 관련된 특징은 크게 두 가지로 나뉠 수 있다. 첫 번째 특징은 동일한 사용자가 짧은 시간 간격으로 연속적으로 게재한다는 것이며, 게재하는 내용은 완전히 동일하거나 일부 단어를 유사어로 치환하기도 한다[11]. 게재 시기와 관련된 두 번째 특징은, 다수의 비정상적 정보가 의견 게재가 가능해진 시점으로부터 짧은 시간 이내에 이루어진다는 것이다 (e.g., 30분~2시간 이내)[12]. 먼저 게재되고 좋은 평가를 받아 목록의 앞을 선점하면, 이후에 작성된 글보다 더 자주 노출되기 때문이다.

(그림 1)은 위에 소개된 특징들을 함께 보여주는 사례이며, 다음 뉴스 사이트 정치섹션8)에서 관찰되었다. 각각의 수직선은 의견이 게재된 시점을 나타내며, 모든 의견은 동일한 사용자에 의해 게재되었다. 의견 게재가 가능해진 시점으로부터 15분 이내에 첫 의견이 게재되었으며, 거의 유사한 내용이 수분 간격으로 3시간 이상 지속적으로 게재되었다.



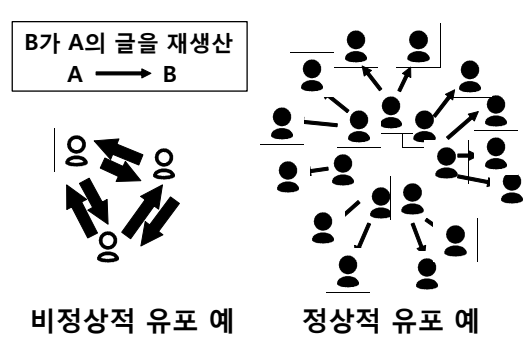
(그림 1) 동일한 사용자가 다수의 의견을 짧은 시간 간격으로 게재한 예

3.2 정보 유포 과정상의 특징 활용

비정상적 정보가 전파되는 형태(Diffusion Pattern)가 정상적인 정보의 형태와 다르다는 점이 탐지에 활용될 수 있으며, 두 가지 특징이 대표적으로 사용된다. 첫 번째 특징은 정보를 재생산하는데 기여한 사용자의 수이다. 재생산이라 함은 트위터의 리트윗처럼 다른 사용자의 글을 인용하여 유포하는 것을 의미한다. 정상 정보에 비해 비정상 정보는 상대적으로 소수의 사용자에 의해 반복적으로 재생산 되는 경향을 보인다[13].

전파 형태와 관련된 두 번째 특징은 정보를 최초로 유입한 사용자의 수이다. 최초 유입이라 함은 해당 정보가 리트윗되기 전 처음 언급된 경우를 의미한다. 첫 번째 특징과 유사하게, 비정상 정보는 상대적으로 소수의 사용자에 의해 최초 언급되고 있다.

(그림 2)는 위의 두 가지 특징을 가진 사례를 보여준다. 화살표가 의견의 재생산을 나타내며, 화살표의 두께는 재생산 횟수를 나타낸다. 정상 정보의 경우 다수의 사용자에 의해 최초 유입된 정보가 다른 다수의 사용자에 의해 재생산되며 확산되어 간다. 비정상 정보의 경우 소수의 사용자에 의해 최초로 유입된 후, 다른 유입자에 의해 반복적으로 재생산되고 있다. 이는 리트윗 횟수

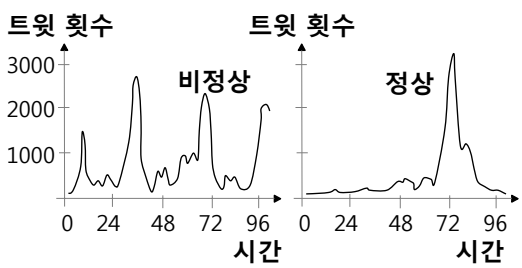


(그림 2) 정보 유포 형태의 차이

8) http://media.daum.net/politics/

를 인위적으로 증가시킴으로써 노출 가능성을 높이기 위함으로 보인다.

전파 형태에서 나타나는 비정상적인 특징은 시간에 따라 게재되는 의견수의 변화에서도 볼 수 있다 (그림 3). 정상 정보의 경우 한두 번 정도의 정점(peak)을 보인 후 서서히 소멸된다. 이에 반해 비정상 정보는 상대적으로 많은 횟수의 정점이 주기를 갖고 나타나며, 이는 인위적으로 전파하려는 노력이 여러 번 반복됨을 보여준다[14].



(그림 3) 시간에 따른 의견수의 변화

3.3 협력자 집단을 함께 탐지

다수의 사용자가 협력하여 함께 비정상 정보를 유포하는 행위는 종종 관찰되는데 이는 개별 사용자가 유포하는 것 보다 효과적이기 때문이다. 따라서 협력이 의심되는 사용자를 함께 관찰하여 탐지하는 기법도 제안되었다. 집단 전체의 행동에서는 개별 사용자보다 비정상적인 특징이 두드러지게 나타나므로 탐지가 더 용이할 수 있다는 장점이 있다.

협력자를 함께 탐지하기 위해서는 먼저 관찰 대상 집단을 선정한다. 주로 n회 이상 동일한 주제에 함께 글을 게재한 사용자들을 협력의 가능성이 있다고 보고 자세한 검사를 수행한다[15]. 특히, 다음과 같은 특징들을 평가한다: 구성원들은 얼마나 유사한 시간대에 활동했는가? 게재한 글의 내용 및 평가 등급은 구성원 간에 서로 유사한가? 게재한 글의 내용 및 평가 등급이 다른 글과 얼마나 차이가 나는가? 이러한 평가의 결과, 만약 집단의 구성원이 짧은 시간 내에 서로 유사한 내용을 게재하였으며, 게재한 내용이나 등급이 다른 사용자들과 크게 차이가 난다면 비정상으로 분류될 가능성이 높아진다.

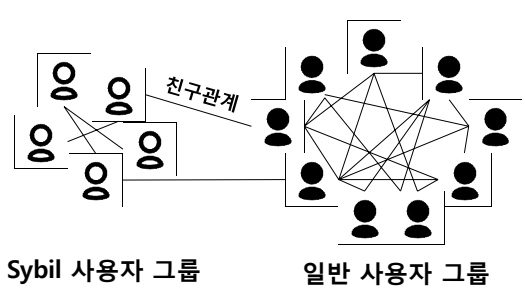
3.4 허위 사용자(Sybil) 탐지

다수의 비정상적 정보는 악성 사용자들이 만든 허위 계정에 의해 게재된다고 알려져 있으며, 이러한 허위 사용자를 Sybil이라 한다. Sybil 탐지방법을 제안한 기존연구의 대부분은 다음과 같은 가정에 기반하고 있다:

Sybil은 일반 사용자들과 친구관계를 맺기 쉽지 않으며, 따라서 일반 사용자 그룹과는 드물게 연결된 Sybil만의 집단을 형성한다[16] (그림 4).

하지만, 이후 연구에서는 Sybil이 일반 사용자들과 조밀한 친구관계를 맺고 있는 경우도 종종 발견되었으며[17], 보수를 받고 친구관계를 임의로 추가해주는 서비스도 존재함이 밝혀졌다[18].

위와 같이 친구관계를 활용하는 대신, 사용자가 웹 사이트에서 수행하는 일련의 이벤트(ClickStream)를 정상적인 경우와 비교하여 Sybil을 탐지하는 기법도 제안되었다[19]. 예를 들어, 정



(그림 4) Sybil 집단과 일반 사용자 집단 간 연결 관계

상 사용자는 사진을 보거나 공유하는 행위를 가장 많이 수행하는데 반해, 비정상 사용자는 글을 게재하거나 친구관계를 맺는 행위의 비중이 훨씬 높았다.

4. 결 론

온라인 공간을 통해 유포되는 비정상적 정보는 사용자들에게 금전적, 정신적 피해를 입힘과 동시에 선거의 결과를 바꿔놓을 수 있을 만큼 그 영향력이 막강하다. 본 논문에서는 이러한 비정상적 정보의 유포를 탐지하여 사전에 차단하기 위한 기법들(주로 게재되는 내용, 시간, 유포하는 사용자의 수 및 이들 간의 온라인상 친구관계에서 발견되는 비정상적인 특징 활용)을 소개하였다.

소개된 특징들을 함께 활용하면, 비정상적 정보를 게재하는 사용자 각각이 다수의 글을 게재하는 경우에는 높은 확률로 탐지 가능하다. 하지만 최근에는 타인명의의 계정을 대량으로 구매하거나 도용하여 다수의 계정을 돌아가며 사용하는 경우가 점차 증가하고 있다. 이러한 경우 개별적인 계정에는 비정상적인 특징이 충분히 나타나지 않아 탐지되지 않을 가능성이 높아진다. 따라서 이에 대비한 개선된 탐지방법의 개발이 필요하다.

참 고 문 헌

[1] M. Frazer and S. Dutta, “Throwing sheep in the boardroom: how online social networking will transform your life, work and world,” Wiley, 2008.

[2] R. M. Bond, C. Fariss, J. Jones, A. Kramer, C. Marlow, J. Settle, and J. Fowler, “A 61-million person experiment in social influence and political mobilization,” in Nature, vol. 489, no. 7415, pp. 295-298, Sep. 2012.

[3] D. Mocanu, L. Rossi, Q. Zhang, M. Karsai, and W. Quattrociocchi, “Collective attention in the age of misinformation,” Computing Research Repository (CoRR), Mar. 2014.

[4] C. Vega, “Yelp outs companies that pay for positive reviews,” ABC news, Nov. 2012, http://abcnews.go.com/blogs/business/2012/11/yelp-outs-companies-that-pay-for-positive-reviews/

[5] C. Sorrel, “Apple expels 1,000 apps after store scam,” CNN news, Dec. 2009, http://edition.cnn.com/2009/TECH/12/09/wired.apple.apps/index.html

[6] “Russian Twitter political protests swamped by spam,” BBC news, Dec. 2011, http://www.bbc.com/news/technology-16108876

[7] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, “Detecting spammers on Twitter,” *in Proc. of CEAS*, 2010.

[8] Y. Freund and R. E. Schapire, “A short introduction to Boosting,” in Journal of Japanese Society for Artificial Intelligence, vol. 14, no. 5, pp. 771-780, Sep. 1999.

[9] Boostrap Aggregating (Bagging) http://en.wikipedia.org/wiki/Bootstrap_aggregating

[10] E. Lim, V. Nguyen, N. Jindal, B. Liu, and H. W. Lauw, “Detecting product review spammers using rating behaviors,” *in Proc. of ACM CIKM*, 2010.

[11] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao, “Detecting and Characterizing Social Spam Campaigns,” in *Proc. of ACM IMC*, 2010.

[12] N. Jindal and B. Liu, “Opinion spam and analysis,” *in Proc. of WSDM*, 2008.

[13] J. Ratkiewicz, M. D. Conover, M. Meiss, B. Goncalves, A. Flammini, and F. Menczer, “Detecting and tracking political abuse in social media,” *in Proc. of ICWSM*, 2011.

[14] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, “Prominent features of rumor

propagation in online social media," *in Proc. of ICDM*, 2013.

[15] A. Mukherjee, B. Liu, and N. Glance, "Spotting fake reviewer groups in consumer reviews," *in Proc. of ACM WWW*, 2012.

[16] L. Alvisi, A. Clement, A. Epasto, S. Lattanzi, and A. Panconesi, "SoK: the evolution of Sybil defense via social networks," *in Proc. of IEEE Symposium on Security and Privacy*, 2013.

[17] Z. Yang, C. Wilson, X. Wang, T. Gao, B. Y. Zhao, and Y. Dai, "Uncovering social network Sybils in the wild," *in Proc. of IMC*, 2011.

[18] M. Motoyama, D. McCoy, K. Levchenko, S. Savage, and G. M. Voelker, "Dirty jobs: the role of freelancer labor in web service abuse," in *Proc. of USENIX Security Symposium*, 2011.

[19] G. Wang, T. Konolige, C. Wilson, H. Zheng, and B. Y. Zhao, "You are how you click: ClickStream analysis for Sybil detection," in *Proc. of USENIX Security Symposium*, 2013.

저 자 약 력

이 시 형

이메일 : sihyunglee@swu.ac.kr

- 2010년 5월 Carnegie Mellon University 전자컴퓨터 공학과 졸업 (공학박사)
- 2010년 7월~2011년 8월 IBM TJ Watson 연구소 (박사후연구원)
- 2011년 9월~현재 서울여자대학교 정보보호학과 조교수
- 관심분야 : 네트워크 관리 및 보안, 의견 분석