KIPS Tr. Software and Data Eng.
Vol.4, No.3 pp.129~134 pISSN: 2287-5905

http://dx.doi.org/10.3745/KTSDE.2015.4.3.129

Unsupervised Scheme for Reverse Social Engineering Detection in Online Social Networks

Hayoung Oh†

ABSTRACT

Since automatic social engineering based spam attacks induce for users to click or receive the short message service (SMS), e-mail, site address and make a relationship with an unknown friend, it is very easy for them to active in online social networks. The previous spam detection schemes only apply manual filtering of the system managers or labeling classifications regardless of the features of social networks. In this paper, we propose the spam detection metric after reflecting on a couple of features of social networks followed by analysis of real social network data set, Twitter spam. In addition, we provide the online social networks based unsupervised scheme for automated social engineering spam with self organizing map (SOM). Through the performance evaluation, we show the detection accuracy up to 90% and the possibility of real time training for the spam detection without the manager.

Keywords : Online Social Networks, Reverse Social Engineering, Unsupervised Scheme

온라인 소셜 네트워크에서 역 사회공학 탐지를 위한 비지도학습 기법

오 하 영†

요 약

역 사회공학 기반 스팸공격은 공격자가 직접적인 공격을 수행하는 것이 아니라 피해자가 문제 있는 사이트 주소, 문자, 이메일 수신 및 친구 수락 등을 통해 유도하기 때문에 온라인 소셜 네트워크에서 활성화되기 쉽다. 스팸 탐지 관련 기존 연구들은 소셜 네트워크 특성을 반영하지 않은 채, 관리자의 수동적인 판단 및 라벨링을 바탕으로 스팸을 정상 데이터와 구분하는 단계에 머물러있다. 본 논문에서는 소셜 네트워크 데이터 중 하나인 Twitter spam데이터 셋을 실제로 분석하고 소셜 네트워크에서 다양한 속성들을 반영하여 정상 (ham)과 비정상 (spam)을 구분할 수 있는 탐지 메트릭을 제안한다. 또한, 관리자의 관여 없이도 실시간 및 점진적으로 스팸의 특성을 학습하여 새로운 스팸에 대해서도 탐지할 수 있는 비지도 학습 기법(unsupervised scheme)을 제안한다. 실험 결과, 제안하는 기법은 90% 이상의 정확도로 정상과 스팸을 구별했고 실시간 및 점진적 학습 결과도 정확함을 보였다.

키워드 : 온라인 소셜 네트워크, 역 사회공학, 비지도학습 기법

1. 서 론

사회공학(Social Engineering: SE)은 인간 상호 작용의 깊은 신뢰를 바탕으로 사람들을 속여 정상 보안 절차를 깨트리기 위한 비기술적 침입 수단으로 기존의 시스템 공격들과 다르다. 예를 들어, 능동적 사회공학 공격자(active social engineering)는 통신망 보안 정보에 접근 권한이 있는 담당자와 신뢰를 쌓고 전화나 이메일을 통해 그들의 약점을 도청하거나 도움을 역으로 이용한다. 그 결과 궁극적으로 시스템 접근 코드와 비밀번호를 알아내 시스템에 침입한다.

반면, 자동화 사회공학(Automated Social Engineering: ASE)에서는 공격자가 직접 공격을 수행하는 것이 아니라 취약한 담당자가 문제 있는 이메일(spamming)을 수신하거나 특정 사이트 주소를 클릭 하도록 유도해 공격을 시작한다. 즉, ASE는 공격을 당하는 희생자가 이메일, 문자 및 사이트 클릭 등을 통해 공격당하기 때문에 역 사회공학(Reverse Social Engineering: RSE)으로도 불린다.

온라인 소셜 네트워크 환경에서는 사용자들끼리 친구 추천, 친구 수락 및 댓글 달기 등이 자유롭기 때문에 자동화 사회공학에 노출되기 쉽다[1-2]. 예를 들어, 트위터(Twitter)는 follower 및 followee와 같은 자유로운 관계로 인해 영향력 있

※ 이 논문은 2014년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(No. 한국연구재단에서 부여한 과제번호: NRF-2014R1A1A1003562)의 일환으로 수행하였음.
† 정 회 원 : 숭실대학교 정보통신전자공학부 조교수
Manuscript Received : October 29, 2014
First Revision : December 19, 2014
Accepted : December 20, 2014
* Corresponding Author : Hayoung Oh(hyoh@ssu.ac.kr)

는 위치(Hubness)가 쉽게 노출되는 구조이다. 많은 followers들은 유명 인사 followee에 노출되어있는 사이트 링크를 클릭, 메시지 수신 및 댓글 달기 등이 순조롭기 때문에 역 사회공학으로 인한 공격에 쉽게 감염(spam)된다. 또한, [1]의 실험에서는 facebook의 최대 41% 사용자들이 사이트에서 랜덤으로 추천해주는 친구를 수락한다는 결과를 보여줬다. 그러므로 정상 사용자가 역 사회공학 기반 공격자와 친구가 맺어지면 보안에 취약한 이메일, 쪽지, 문자 및 사이트 댓글 등도 수락하거나 클릭하기 쉽다.

[3]의 Jagatic et al. 저자들은 온라인 소셜 네트워크 정보 기반 이메일을 통해 개인 정보를 알아내어 그들의 돈을 빼돌리는 피싱 사기(phishing)의 심각성을 보여줬다.

[4]에서는 Bayesian 기법을 활용하여 단문메시지서비스 스팸(Short Message Service spam: SMS spam)을 필터링하는 데 초점을 두었다. SMS spam을 표현하기 위해서 기존의 spam email 방식을 참고하여 영어와 스페인어로 구성된 방대한 실험 데이터를 만들었다. [5]에서는 [4]의 연구를 확장하여 spam email를 구분하는 데 사용된 내용어 기반 분류기법(content filtering)을 그대로 SMS spam을 구분하는 데 활용했다. 이를 위해 메시지를 앞부분(head), 본문(text) 및 기타 부분으로 분리하고 관리자의 수동적인 해석이 필요한 데이터 마이닝 분류 기법들을 활용하여 유사한 특성대로 분류하고 중요 요소(features)들을 분류했다.

하지만 spam 관련 기존 연구들은 키워드 분석 기반 필터링 및 특정 계정별 정보 수신 비율 등에 초점을 두고 해결책을 제시했기 때문에, 발신 계정 도용(account spoofing 혹은 Sybils) 등과 같이 공격타입이 좀 지능화되거나, 소셜 네트워크 개념을 활용한 온라인 공격은 해결할 수 없다. 또한 사회공학 공격 문제는 Twitter, Facebook 등과 같은 소셜 네트워크 혹은 이메일 서비스에 한정되어 관리자가 수동으로 참여해야 하는 지도학습 연구(supervised scheme)가 진행되었고 실시간 비지도학습 기법(unsupervised scheme)에 접목하지 못했다.

즉, 초기 스팸 탐지 기법들은 이미 알려진 스팸에 대한 정보를 수동적으로 시스템에 인코딩 하여 스팸 여부를 판단하는 방법으로 규칙의 생성 및 확장이 매우 어렵고 그 효율성도 매우 떨어지는 방법이다. 따라서 소셜 네트워크 특성을 고려하여 인공지능, 기계 학습 및 데이터 마이닝 기법들을 스팸 탐지에 적용하는 연구가 늘어나는 추세이나 아직까지 많은 연구가 여전히 분류(classification) 방법을 포함한 지도학습 알고리즘에 근간을 두고 있어 다음과 같은 문제점들이 있다. 먼저 소셜 네트워크 데이터 학습과 스팸 탐지 과정이 확연히 구분되어있고 탐지 과정 전에 충분한 학습 과정이 이루어져야 하므로 안정적인 성능이 나오기까지 많은 비용이 든다. 그리고 학습을 위해 많은 양의 분류된 데이터(labeled data)를 필요로 하는데 이러한 방대한 양의 학습 데이터를 수집하고 분류하는 것은 매우 어려운 일이며 학습 데이터의 질에 의해 탐지 성능이 크게 좌우된다[7-8]. 또한 실시간으로 스팸 성격을 반영하는 점진적 학습의 수행이 불가능하고 학습된 데이터 이외의 새로운 스팸 유형에 대한 탐지가 어렵다. 대안으로 비지도 학습(unsupervised learning)의 데이터 마이닝 기법을 이용한 연구가 진행되었으나, 비지도 학습만 사용했을 경우에는 학습 시 입력 데이터에 대해 어떠한 정보도 주지 않으므로 그 결과에 대한 해석이 힘들다는 문제점이 발생한다.

이 모든 것을 고려하여 본 논문에서는 온라인 소셜 네트워크에서 역 사회공학으로 인한 스팸 탐지 기법을 처음으로 제안한다. 이를 위해, 비지도 학습인 자기 조직화 지도(Self Organizing Map: SOM)와 지도 학습으로 소셜 네트워크 속성들(i.e., 중간 중심성(Betweenness Centrality), 클러스터링 계수(Clustering Coefficient)) 등을 결합하여 점진적 학습 및 실시간 스팸 탐지가 가능한 SOM 기반 스팸 속성 상관관계 메커니즘을 설계하고 이를 실시간 스팸 탐지에 활용한다. 즉 비지도 학습 SOM은 점진적 학습과 실시간 탐지가 가능하지만, 학습 결과 지도 해석이 힘들다는 문제점이 있기 때문에 이를 해결하기 위해 분류되어있는 Twitter spam 데이터[9]를 사용하여 속성 간의 상관관계를 분석하여 규칙을 생성하고 이를 기반으로 결과에 대한 정보를 알아낸다.

본 논문의 구성은 다음과 같다. 1절의 서론에 이어서 2절에서는 시스템 모델을 살펴본다. 3절에서는 실시간 스팸 탐지를 위한 SOM 기반 스팸 속성 상관관계 메커니즘을 제안하고, 4절에서는 제안한 스팸 탐지 메커니즘을 다양한 측면에서 실험한 내용과 결과를 기술한다. 마지막으로 5절에서는 본 연구의 결론 및 향후 연구 계획에 대해서 기술한다.

2. 시스템 모델

2.1 Twitter spam 데이터 기반 소셜 네트워크의 속성들

David Easley와 Jon Kleinberg는 [11]의 저서에서 소셜 네트워크의 다양한 속성에 대해서 정의했다. 본 연구에서는 Twitter spam 데이터 셋에서 아래와 같은 소셜 네트워크 속성들을 재정의하여 활용한다.

1) 중간 중심성(Betweenness Centrality)

중간 중심성(Betweenness Centrality: BC)은 노드가 얼마나 주요 길목에 위치하는지를 표현해줄 수 있는 척도이다. 특정 노드 n의 BC값은 모든 노드 쌍들에 대해서 그 둘을 잇는 모든 최단 경로들(shortest paths) 중에 특정 노드 n을 지나는 가장 짧은 최단 경로 비율로 Equation (1)처럼 계산 된다[11].

$$BC(n) = \frac{1}{|V|(|V|-1)} \sum_{s,t \in V} \frac{\sigma(s,t|n)}{\sigma(s,t)} \quad (1)$$

$\sigma(s,t)$는 노드 s와 t 사이의 최단거리 경로의 개수, $\sigma(s,t|n)$는 노드 s와 t 사이의 최단거리 경로의 개수 중 노드 n을 지나는 경로들의 개수이다. 그리고 V는 존재하는 모든 노드의 집합을 의미한다. Fig. 1은 10개의 노드가 2가지 시나리오로 존재할 때 각 노드의 BC값을 보여준다. Fig.

1(a)는 각 노드가 이웃 노드들과만 1홉 연결선이 존재하는 소셜 네트워크를, Fig. 1(b)는 각 노드가 이웃 노드들뿐만 아니라 최대 3홉까지 연결선을 가질 수 있는 소셜 네트워크를 보여준다. 그 결과, 각 노드가 전체 토폴로지에서 중앙에 위치하거나, 다양한 이웃들과 연결선을 많이 유지할수록 해당 노드의 BC값이 커짐을 알 수 있다. 본 연구에서는 소셜 공학으로 인한 followee들의 중간 중심성을 각각 계산하여 경향을 파악한다.

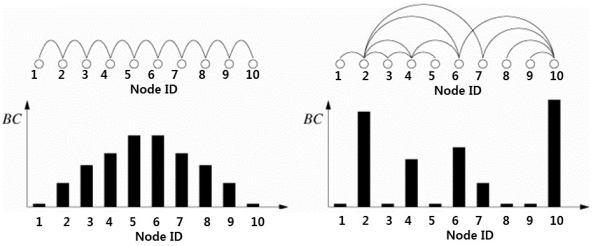

Fig. 1. Betweenness Centrality

2) 클러스터링 계수(Clustering Coefficient)

클러스터링 계수(Clustering Coefficient: CC)는 특정 노드와 그 이웃한 노드들 사이에서 생길 수 있는 모든 가능한 연결선들의 개수 중에 실제로 존재하는 연결선의 개수를 의미한다. Equation (2)는 노드 n의 CC를 계산하는 계산식을 보여준다[11].

$$CC(n) = \frac{2e_n}{deg(n)(deg(n)-1)} \quad (2)$$

e_n은 노드 n의 이웃들 간의 실제 연결선의 수를 의미한다. deg(n)은 노드 n이 얼마나 많은 연결선을 가지고 있는지 정도(degree)를, 즉 이웃한 노드의 개수를 의미한다. Fig. 2는 특정 노드 n에서 이웃한 노드들 {1,2,3} 사이에서 실제로 존재하는 연결선의 개수를 고려하여 특정 노드 n의 CC 값을 구한 예시를 보여준다. 소셜 네트워크에서는 노드 n 그 자체도 중요하지만 노드 n의 이웃 노드(i.e., facebook에서는 친구)들 간의 관계도 고려해야 하기 때문에 CC는 중요한 평가요소 중 하나이다. 본 연구에서는 소셜 공학으로 인한 followee들을 follow하고 있는 followers의 클러스터링 계수를 각각 계산하여 경향을 파악한다.

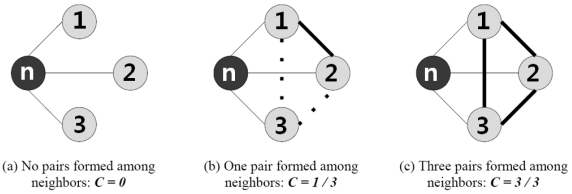

Fig. 2. CC of a Specific Node n

3) 동종애의 원칙(Homophily principle)

동종애의 원칙은 소셜 네트워크에서 사회적 지위나 직업, 성향이 비슷할수록 사람들은 서로 친근감을 느끼게 되고, 상대적으로 많이 상호작용하며 긴밀한 네트워크를 구성한다는 것이다. 시간이 지나면서 네트워크는 제도화되고 내부 커뮤니케이션을 강화하여 "동질성"을 더욱 증가시키게 된다.

[10]에서는 Twitter네트워크에서 의심이 되는 followers들과 그들이 follow하고 있는 followee들을 찾아서 그룹화해보면 그들끼리 동질성(Synchronized Behavior)은 높고 정규화(Normality)는 낮게 만들어진다는 것을 실제 데이터 기반 실험을 통해 증명했다. 즉, 이 논문에서는 spammer로 의심되는 follower 혹은 followee들이 비슷한 개수의 in-degree 혹은 out-degree를 가지면서 비슷한 수치의 HITS value(Hubness or Authoritativeness)를 갖고 있다는 것을 밝혔다. Synchronicity는 특정 소스 노드 follower의 몇몇 followee들을 대상으로 그들을 비교한 평균 유사도(closeness)를 계산한 것이고, Normality는 특정 소스 노드 follower의 몇몇 followee들과 전체 노드들을 비교한 평균 유사도를 계산한 것이다. 그 결과, 의심스러운 spam계정들은 특정 followee를 공격하기 때문에 높은 synchronicity를 갖지만 전체적인 경향과 다르기 때문에 낮은 normality를 갖는다.

[13]에서는 facebook과 instagram 데이터 셋을 처음으로 응용서비스(applications)별로 구분하고 거짓 계정(fake accounts) 및 망가진 사용자 계정(compromised user accounts)을 분석했다. 이를 위해, 일정 시간 간격(sliding window) 동안 특성이 비슷한 계정들끼리 가중치를 높게 주고 유사한 특성을 보이는 계정들을 그룹화 했으며 방대한 데이터 셋을 효율적으로 처리할 수 있는 분산처리 기법도 추가로 제안했다. 이 논문에서는 매우 짧은 기간 동안에도 1156개의 의심되는 클러스터링 되는 그룹(campaign)들이 발견되고 그 안에 200만 개 이상의 거짓 계정들이 있음을 밝혔다.

하지만 [10]에서는 spam으로 이미 라벨링 된 Twitter spam 데이터 셋에서 spammer의 패턴을 밝히고 라벨링 되어있지 않은 새로운 Twitter데이터에 대한 해결책은 제시하지 못했다. 또한 동적으로 발생하는 새로운 소셜 네트워크 Twitter계정에 대해서 점진적으로 훈련하여 기존의 라벨링 된 spammer들과 어떤 연관성이 있는지 밝히지 못했다[10, 13].

2.2 역 사회공학에 대응하기 위한 탐지 메트릭

본 논문에서는 [10]와 같이 spam follower들이 유명 인사 등과 같은 특정 몇몇 followee에 공격 목표(target)을 둔다고 가정했다. 또한, spam follower들은 동일한 목표를 가지고 있기 때문에 비슷한 위치에 몰려있고 몇몇은 정상 데이터 쪽과 연결선을 가지고 있다는 것을 데이터 기반 실험 및 [12] 참고문헌의 연구 결과를 바탕으로 가정했다. 이를 증명하기 위해 Twitter Accounts의 followee들(target)의 동질성(Synchronicity), 중간 중심성(Betweenness Centrality) 및 클러스터링 계수(Clustering Coefficient)의 평균을 in-degree와 authoritativeness로 만든 2D 그래프에 표현한다. 각 followee들의 위치를 표현하기 위해 그래프는 2의 거듭제곱(powers of 2) 단위로 작게 구분한다. 그 결과, 많은 연결선

을 가진 몇몇의 followee들과 적은 연결선을 가진 많은 followee들로 구성된 power-law 분포에서 spam followee들이 특정 위치에 몰려있음을 알 수 있다. 특히, [14]의 기법에서와 같이 일정 시간 간격으로 관찰 결과, 역 소셜 공학 followee들은 영향력을 최대화하기 위해 실제 연결선인 클러스터링 계수가 높은 follower들과 연관되어있고 노드들의 주요길목에 있기 때문에 중간 중심성도 높다는 것을 알 수 있었다.

$$\max \frac{1}{T}\sum_{t=1}^{T}(BC+CC+Sycronicity) \quad (3)$$

Equation (3)은 제안하는 기법을 반영하는 메트릭으로 t는 해당 특성을 관찰하는 시간 간격을 의미한다. 제안하는 기법은 소셜 네트워크 속성들을 고려하여 역 사회공학 spam followee들의 위치 연관성을 살펴보기 때문에 기존 연구[10]보다 역 소셜 공학 대응에 더 적합하다고 볼 수 있다.

3. 자기 조직화지도(SOM) 기반 역 사회공학 스팸 속성 상관관계 메커니즘

실시간 사회공학 spam 탐지를 위해 앞에서 설명한 역 사회공학에 대응하기 위한 메트릭 정보와 비지도 학습의 데이터 마이닝 기법인 SOM을 이용하여 스팸 속성 상관관계를 분석한다[14-15]. 스팸 속성 상관관계는 정상(ham)과 스팸의 성격을 반영하므로 궁극적으로 실시간 스팸 탐지 메커니즘에 사용될 수 있다. 즉, 정상과 스팸은 제안하는 메트릭 수식인 Equation (1)을 기반으로 서로 다른 속성 상관관계를 보이기 때문에 스팸 탐지가 가능하며, 추후 다양한 종류의 비정상 스팸들의 각 특징을 비교해보고 분석해볼 수 있다. SOM 기반 스팸 속성 상관관계분석을 이용한 실시간 스팸 탐지 메커니즘은 Fig. 3처럼 크게 3단계로 이루어진다. 첫 번째는 전처리와 정규화된 실험데이터로 탐지에 필요한 랩을 생성하는 학습 단계, 두 번째는 학습된 맵에서 정상 및 스팸 속성 상관관계를 활용한 각 클러스터별 분류단계, 마지막으로 실시간 탐지와 점진적 학습이 이루어지는 단계이다.

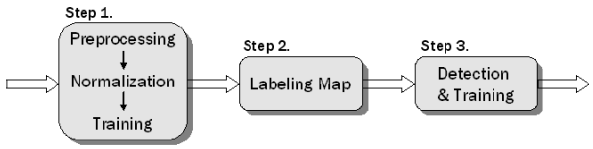

Fig. 3. Real Time Spam Detection Mechanism

3.1 학습과정(Training)

1) 전처리 과정(Preprocessing)

학습 단계 전에 스팸 탐지에 효과적인 정보로 데이터를 변형해주는 전처리 단계가 필요하다. 이 단계에서는 Twitter Accounts의 followee들(target)의 동질성(Synchronicity), 중간 중심성(Betweenness Centrality) 및 클러스터링 계수(Clustering Coefficient)를 계산하는 단계가 포함된다.

2) 정규화 과정(Normalization)

SOM에서는 각 속성 맵들의 형태를 결합하여 모든 속성이 고려된 U-matrix를 최종적으로 생성한다. 그런데 Twitter Accounts의 followee들(target)의 동질성(Synchronicity), 중간 중심성(Betweenness Centrality) 및 클러스터링 계수(Clustering Coefficient) 등 각 속성마다 다양한 범위의 값을 갖는다. 이 상태에서는 다른 속성에 비해 범위가 큰 속성값이 U-matrix를 생성하는 데 많은 영향을 끼치게 되어 모든 속성을 평등하게 고려한 U-matrix가 생성되지 않으므로 모든 속성이 균등하게 U-rnatrix 생성에 반영될 수 있도록 데이터값을 정규화하는 과정이 필요하다.

3) SOM 알고리즘을 이용한 학습 과정(Training)

Kohonen에 의해서 개발된 자기 조직화 지도인 SOM은 신경망 기법을 사용하는 클러스터링의 모델이면서 비지도 학습을 사용한다는 것이 특징이다. 비지도 학습은 지도 학습과 달리, 주어진 입력에 대해 정확한 해답을 주지 않고 자기 스스로 학습을 한다. 스팸 탐지에 있어서는 미리 정상과 비정상으로 분류된 학습 데이터가 필요하지 않고, 분류되지 않은 학습 데이터를 넣어주면 비슷한 성격의 데이터끼리의 클러스터링을 통해 기계 스스로가 정상과 비정상 트래픽으로 분류해준다. 또한 입력 데이터와 가장 가까운 뉴런의 이웃 뉴런들도 비슷한 방향으로 함께 학습시키기 때문에 인접한 뉴런들은 비슷한 성격을 가칠 것이라고 예측할 수 있다. SOM의 학습 알고리즘은 5단계로 이루어지며, 이는 Table 1과 같다.

① Initialize the network
For each node i set the initial weight vector $w_i(0)$ to be random.
Set the initial neighborhood $N_i(0)$ to be large.
② Present the input
Present $x(t)$, the input pattern vector x at time t ($0<t<n$ where n is the number of iterations defined by the user) to all nodes in the network simultaneously.
x may be chosen at random or cyclically from the training data set.
③ Calculate the winning node
Calculate node c with smallest distance between the weight vector and the input vector,
$$\| x(t) - w_c(t) \| = \min_i \{ \| x(t) - w_i(t) \| \}$$
hence
$$c = \arg\min_i \{ \| x(t) - w_i(t) \| \}$$
④ Update the weights
Update weights for c and nodes within neighborhood $Nc(t)$
$$w_i(t+1) = w_i(t) + h_{ci}(t)[x(t) - w_i(t)] \text{ if } i \in N_c(t)$$
$$= w_i(t) \text{ if } i \notin N_c(t)$$
where h_{ci} is a scalar "kernel" function,
$$h_{ci}(t) = \alpha(t) \cdot \exp\left(-\frac{\| x_i - r_c \|^2}{2\sigma(t)^2}\right)$$
⑤ Present the next input
Decrease h_{ci} so that $h_{ci}(t+1) < h_{ci}(t)$
Reduce the neighborhood set so that $N_i(t+1) \subset N_{ci}(t) \ \forall_i$
Repeat from step 2 choosing a new unique input vector $x(t+1) \neq x(j), j \leq t$
until all iterations have been made ($t=n$).

Table 1. SOM Algorithm

3.2 분류 과정(Map Labeling)

학습 단계를 거친 후 생성된 맵은 입력 데이터에 대한 어떠한 정보도 주지 않기 때문에, 사용자는 SOM을 이용해 구분된 클러스터가 정상인 ham 혹은 비정상인 spam인지 분간하기 어렵다. 이를 해결하기 위해 속성별 맵의 유사도를 보고 속성 간의 상관관계를 분석하여 상관계수가 높은 속성 집합을 기반으로 규칙을 생성하여 맵의 클러스터 구분이 가능하게 한다. 두 속성이 어느 정도 연관성이 있는지 상관관계를 분석하기 위해 피어슨(Pearson) 상관계수를 변형하여 사용한다. 속성별 상관관계를 분석하기 위해 Twitter spam 실험 데이터에 포함된 정상 및 스팸별 데이터를 분류하여 각각 속성별 맵을 형성한다.

3.3 실시간 역 사회공학 스팸 탐지 및 점진적 학습과정 (Detection and Training)

학습을 통해 생성된 U-matrix를 이용한 실시간 스팸 탐지는 SOM 알고리즘의 일부분인 유클리디언 거리 측정을 이용하여 이루어진다. 탐지 알고리즘은 Table 2와 같다.

BMU=arg_{min} ‖ ($x(n)$-$w_j(n)$ ‖
where $x(n)$ is the input vector being presented at time n
and $w_j(n)$ is the weight vector for all nodes in the network.
If BMU ∈ the set of *normal clusters*,
then $x(n)$=*normal*
else $x(n)$=*abnormal*

Table 2. Spam Detection Algorithm Based on Euclidean Distance Equation

스팸 탐지 후에 입력 데이터의 순차적인 학습을 통하여 승자 뉴런의 가중치와 인접한 이웃 뉴런의 가중치가 조정되고 시간이 흐름에 따라 맵이 갱신되며 점진적 학습이 이루어지면서 실시간 특성을 반영한 스팸 탐지가 이루어지게 된다. 이는 Table 3과 같다.

$$w_j(n+1)=w_j(n)+\alpha(n)\beta_{i(x)}(n,j)(x(n)-w_j(n))$$

Table 3. Real Time Training Algorithm

4. 성능 평가

4.1 실험 데이터

본 연구에서는 Twitter spam데이터 셋[9]을 분석하고 소셜 네트워크 속성들을 계산하여 활용한다.

4.2 SOM을 제공하는 도구

1) SAS Enterprise Miner

SAS Enterprise Miner는 대부분의 데이터 마이닝 기법들을 제공하는 고차원적인 데이터 마이닝 도구이다. 입력 데이터의 선정(sample), 탐색(explore), 변형(modify), 모델 생성(model), 평가(assess)라는 각 범주별로 여러 가지 방법들을 제공해준다.

본 연구에서는 SOM/Kohonen 노드를 이용하여 실험하였고 사용자는 분석 방법의 설정, 자기 조직화 지도의 크기 설정, 군집수의 지정, 이웃 노드의 학습률, 학습 방법 선택 등의 옵션을 다양하게 지정할 수 있다.

2) Matlab 기반의 SOM Toolbox

SOM Toolbox는 Helsinki University of Technology에서 연구 목적으로 SOM의 쉬운 사용을 위해 개발한 프리웨어의 SOM프로그램 패키지로서, 특히 시각화 측면에서 아주 뛰어나다. 데이터의 전처리, 초기화와 학습, 맵 크기의 설정, 맵의 시각화 및 분석 등 다양한 기법을 제공하며 숫자형의 변수들만 적용된다는 특징이 있다.

Fig. 4는 제안하는 기법으로 정상과 비정상을 탐지하는 비율을 보여준다. 앞서 언급했듯이 역 소셜 공학 기반 비정상 spam followee들은 영향력을 최대화하기 위해 실제 연결선인 클러스터링 계수가 높은 follower들과 연관되어있고 노드들의 주요길목에 있기 때문에 중간 중심성이 높다. 또한, 이미 연결선이 많은 followers들을 간헐적으로 연결하여 spam화하기 때문에 클러스터링 계수가 높으며 같은 입력 연결선 및 정상데이터들과 다른 위치에 그룹화 되는 경향이 있기 때문에 동질성은 낮고 정규화에서는 멀다. 그 결과, Equation (3)으로 정의됐던 스팸 탐지 메트릭 평균값이 spam의 특정구간에서 매우 높게 나타남을 알 수 있다. 제안하는 기법의 탐지 정확도를 검증하기 위해서 수동적으로 라벨링 된 결과와 비교 결과의 탐지율을 Equation (4)를 바탕으로 계산한 결과 10% 미만인 걸로 보아, 제안하는 기법의 탐지 성능은 거의 90% 이상 정확하다고 볼 수 있다.

$$\text{탐지율}=\frac{\text{비정상으로 정확히 판정된 비정상 데이터의 개수}}{\text{전체 비정상 데이터 개수}}\times 100 \quad (4)$$

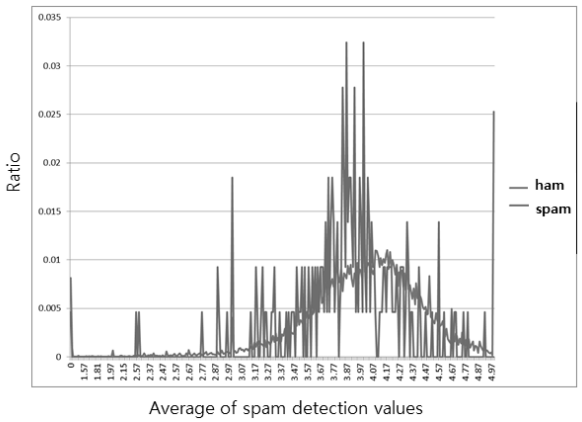

Fig. 4. Average of Spam Detection Values

5. 결 론

본 논문에서는 Twitter spam 실제 데이터를 분석하고 온

라인 소셜 네트워크의 속성들을 재정의하여 비정상 스팸 followee들을 탐지하는 메트릭을 제안했다. 메트릭을 통해 소셜 네트워크에서 역 사회망 공학으로 인한 비정상 스팸을 정확도 높게 탐지했다. 또한 관리자의 수동적인 라벨링이 필요 없이 실시간 및 점진적으로 학습하여 새로운 스팸도 탐지할 수 있는 비지도 학습 기법을 최종적으로 제안했다. 이를 위해 비지도 탐지 알고리즘 중 하나인 자기 조직화 기법을 활용했다. 실험 결과, 제안하는 기법은 비정상 스팸 데이터와 정상 데이터를 확연히 구분하고 실시간 및 점진적 학습 결과도 거의 정확함을 알 수 있었다.

본 연구는 다양하고 방대한 소셜 네트워크 데이터 셋을 정상과 의심되는 그룹으로 구별할 수 있는 기초 연구가 될 수 있으며 추후 사용자 계정 및 응용별 세분화해서 좀 더 심화된 공격 종류들까지 구분할 때 활용될 수 있을 것이다.

References

[1] Sophos Facebook ID Probe. http://www.sophos.com/pressoffice/news/articles/2007/08/facebook.html, 2008.

[2] D. Irani, M. Balduzzi, D. Balzarotti, E. Kirda, and C. Pu, “Reverse social engineering attacks in online social networks,” in *Detection of Intrusions and Malware, and Vulnerability Assessment*, ed: Springer, pp.55-74, 2011.

[3] Jagatic, T. N., Johnson, N. A., Jakobsson, M., and Menczer, F. Social phishing. *Commun. ACM*, Vol.50, No.10, pp.94-100, 2007.

[4] J. M. Gómez Hidalgo, G. C. Bringas, E. P. Sánz, and F. C. García, “Content based SMS spam filtering,” *in Proceedings of the 2006 ACM symposium on Document engineering*, pp.107-114, 2006.

[5] G. V. Cormack, J. M. Gómez Hidalgo, and E. P. Sánz, “Spam filtering for short messages,” in Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, pp.313-320, 2007.

[6] Liu JY, Zhao YH, and Zhang ZX et al. “Spam short messages detection via mining social netwoks,” JOURNAL *OF COMPUTER SCIENCE AND TECHNOLOGY*, Vol.27, No.3, pp.506-514, May, 2012. DOI 10.1007/s11390-012-1239-7

[7] Richard Bassett et al., “DATA MINING AND SOCIAL NETWORKING SITES: PROTECTING BUSINESS INFRASTRUCTURE AND BEYOND,” *Issues in Information Systems*, Vol.XI, No.1, 2010.

[8] Mariam Adedoyin-Olowe, Mohamed Medhat Gaber, and Frederic Stahl, “A Survey of Data Mining Techniques for Social Network Analysis,” Cornell University.

[9] Kurt Thomas, Chris Grier, Vern Paxson, and Dawn Song, “Suspended Accounts in Retrospect: An Analysis of Twitter Spam,” Internet Measurement Conference(IMC), 2011.

[10] Meng Jiang, Peng Cui, Alex Beutel, Christos Faloutsos, and Shiqiang Yang, “CatchSync : Catching Synchronized Behavior in Large Directed Graphs,” KDD ‘14

[11] David Easley, Jon Kleinberg, “Networks, Crowds, and Markets: Reasoning About a Highly Connected World,” *Cambridge University Press*.

[12] Neil Zhenqiang Gong, Mario Frank, and Prateek Mittal, “SybilBelief: A Semi-supervised Learning Approach for Structure-based Sybil Detection,” *IEEE Transactions on Information Forensics and Security*, Vol.9, No.6, 2014.

[13] Qiang Cao, Xiaowei Yang, Jieqi Yu, and Christopher Palow, “Uncovering Large Groups of Active Malicious Accounts in Online Social Networks,” ACM CCS 2014

[14] Hayoung Oh, Jiyoung Lim, Kijoon Chae and Jungchan Nah, "Home Gateway with Automated Real-Time Intrusion Detection for Secure Home Networks," Computational Science and Its Application-ICCSA 2006 Lecture Notes in Computer Science, Vol.3983, pp.440-447, 2006.

[15] Kyoungae Hwang, Hayoung Oh, Jiyoung Lim, Kijoon Chae, and Jungchan Nah, “Traffic Attributes Correlation Mechanism based on Self-Organizing Maps for Real-Time Intrusion Detection,” *Information Processing Society Journal*, Oct., 2005.

오 하 영

e-mail : hyoh@ssu.ac.kr

1998년～2002년 덕성여자대학교

2001년～2004년 신한금융지주회사 e-신한

2004년～2006년 이화여자대학교 컴퓨터공학 (석사)

2006년～2013년 서울대학교 컴퓨터공학 (박사)

2010년 4월～2010년 10월 U.C. Berkeley 방문연구원

2013년 3월～2013년 8월 서울시립대학교 연구교수

현 재 숭실대학교 정보통신전자공학부 조교수

관심분야 : 소셜 정보망, 추천시스템, 무선 네트워크 및 비디오 스트리밍