사이버 범죄 정보 공유를 위한 SNS 빅데이터 활용

신예진*, 오지선*, 이해은*, 함나연*, 조윤진**
*서울여자대학교 정보보호학과, **서울여자대학교 경영학과

Application of SNS Big data in Sharing system of Cyber security crime

Shin, Ye Jin˚, O, Ji Seon˚, Lee, Hae Eun˚, Ham, Na Youn˚, Cho, Yun Jin˚
Seoul Women's University
E-mail : nsu012@gmail.com, jiseon41257@gmail.com , ihaeeun16@gmail.com, hamny888@naver.com, choyj0517@gmail.com

요 약

사이버 범죄는 나날이 증가하고 있는 반면, 일반인을 대상으로 하는 위협 정보 제공은 미비하다는 단점이 있다. 이 점을 해소하고자 개인이나 일반 사업체는 트위터나 SNS를 통해서 정보를 접할 수 있는데, 이런 방대한 데이터에서 개인에게 필요한 정보만 정제되는 프로그램이 없어 접근성이 어렵다는 문제가 있다. 본 논문에서는 개인에게 필요한 보안 정보를 제공하고자 침해 정보를 정제한 서비스에 대해 제공하고 있다.

1. 서론

경찰청의 2017 사이버 위협 보고서에 따르면 2016년 대비 2017년에 사이버 범죄가2018 사이버 위협 보고서에 따르면 2017년에 비해 2018년에 사이버범죄가 13.6% 증가했다.

경찰청 사이버범죄 통계자료에 따르면 해킹, 서비스 거부공격, 악성프로그램 등에 의한 정보통신망 침해형 범죄는 2016년 2,770건, 2017년 3,156건, 2018년 2,888건으로 해마다 약 3,000여건 가까이 발생한다. 사이버 범죄에 대응하기 위해 공격자 IP, URL, 파일의 해시 정보 등의 빅데이터는 중요한 자원이다. 국내에서는 한국인터넷진흥원과 침해사고 대응 기관, 보안 업체들이 이러한 정보들을 공유하고 있다. 그러나 현재의 사이버 위협 정보 공유 체계는 한국인터넷진흥원의 일방향적 정보 공유 성격이 강하고, 공유의 대상이 침해사고 대응 기관, 보안 업체들이기 때문에 개인이나 일반 사업체들에서는 위협 정보를 얻기 어렵다.

개인이나 일반 사업체는 트위터와 같은 SNS를 이용해 빠르게 위협 정보를 얻을 수 있다. 트위터의 CVEnew 계정은 '정보보안 취약점 표준 코드 CVE(Common Vulnerabilities and Exposures)' 정보를 게시하고, IpNigh 계정은 Phishing 공격의 도메인 정보를 공유한다. 하지만 트위터에 게시된 텍스트 중 필요한 정보만 정제하는 별도의 프로그램이 없다면 수많은 양의 정보들을 신속하게 활용하기 어렵다.

본 논문에서는 트위터에서 공유되는 사이버 범죄 관련 정보들을 크롤링하고, 이를 정제하여 개인이

나 사업체에 신속하게 제공하는 서비스에 대해 제안한다.

2. 본론

2.1 Crawling

사이버 범죄가 발생하기 전, 위협 정보들을 미리 알려주는 트위터들이 존재한다. 이러한 트위터의 내용을 수집하여, 해당 사이트나 기업에 알려 사전 대응이 가능하게 한다.

트위터를 크롤링하는 방법에는 트위터에서 공식적으로 제공하는 Tweepy API를 이용하는 방법과 python library인 getOldTweets를 이용하는 방법이 있다. 본 논문에서는 이 중 전자인 tweepy를 이용하고자 한다.

위협 정보에 관한 트윗을 수집하기 위해서는 '취약점', '악성코드', '해커', '위협정보', '보안' 등의 수집할 키워드를 정한다. 정리된 키워드를 기반으로 위협 정보 알림 트윗들을 크롤링한다. 키워드 별로 크롤링한 트윗에서 중복된 내용의 트윗은 제거한다. 중복 트윗 제거 과정 이후, 리트윗이나 광고와 같은 의미 없는 트윗들 또한 제거한다. 정리된 트윗들 중 공격자 IP, URL, 파일의 해시 정보 값이 포함된 트윗을 추출한다.

2.2 Parsing

해당 정보들을 추출하기 위해서 파싱 과정이 필요한데, 이 때는 주로 정규표현식을 이용한 데이터 정제 과정을 거치게 된다. IP, URL, 파일 해시 정보 등의 패턴을 파악하여 해당 패턴을 보이고 있는 문자열을 추출한다. 이렇게 추출해 낸 문자열은 바이러스 토탈로 보내져 해당 정보가 실제 악성 정보인지 검증 과정을 거친다. 추후 파싱 과정에서 정규표현식을 넘어 자연어처리를 이용하도록 머신러닝을 적용하게 된다면 정규표현식이 갖는 오탐의 한계를 극복하고 더욱 정확한 결과를 확보할 수 있을 것으로 기대된다.

2.3 서비스 제공

본 논문에서 제안하는 서비스는 최종적으로 웹의 형태로 사용자에게 제공되게 기획되었다. 웹의 사용빈도와 편리성을 고려하였을 때, 접근성이 뛰어나기 때문에 웹의 형태로 침해정보를 제공하게 된다. 크롤링을 통해 수집된 트윗들의 정보는 Virus Total에서 한번 더 검증을 거치게 되고, 정확한 정보만을 추출하여 웹으로 제공되게 된다.

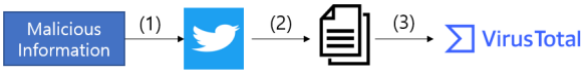

그림 1 서비스 흐름 구조도

본 서비스의 구조도는 다음과 같다.

(1)화이트해커들이 수집하여 트위터에 업로드 하는 악성 정보들 중에서 특정 키워드와 일치하는 정보를 크롤링의 범위로 정한뒤, (2) 크롤링을 통해 수집한다. (3)수집된 정보들은 바이러스 토탈에 자동화 업로드 하여 한번 더 검증을 거쳐 정확한 정보를 산출하게 된다.

산출된 정보들은 편리하게 제공되기 위해 반응형 웹을 적용하여 접근성을 높이고, 새로운 정보가 업로드 될 때마다 사용자에게 알림을 주는 기능을 구현하도록 기획되었다. 웹의 형태를 통해 사용자들은 포털사이트에서 기사를 검색하고 읽는 것 처럼, 최신 공격과 대응에 대한 정보를 발빠르게 접할 수 있다.

3. 결론

본 논문은 사이버 범죄 정보 공유를 위한 서비스 제안으로, SNS상에서 공유되는 위협 정보 데이터들을 수집하는 크롤링 단계, 수집한 정보들을 정규표현식을 이용하여 정제하는 파싱단계, Virus Total 검증과정을 거쳐 최종적으로 사용자에게 정보를 제공하는 웹서비스 총 3단계로 이루어져 있다. 본 서비스는 접하기 어려운 사이버 범죄 관련 정보들을 개인이나 사업체에 빠르게 제공함으로써 사이버 범죄 예방과 정보보호에 대한 인식을

제고시킬 수 있을 것으로 기대된다.

[참고문헌]

[1] 박철민; 조정식. 국외 사이버 위협 정보공유의 체계조사. 한국인터넷진흥원 Report, 2014.

[2] https://cyberbureau.police.go.kr/share/sub3.jsp?mid=030300 [2019년 10월 11일]