딥웹 환경에서 사이버범죄 정보 수집분석 구현

황덕현*, 박소영*, 배지선*, 정송주*, 홍진근*, 박현주**

*백석대학교 ICT학부
**(주)시옷 CIOT, Inc

madbrain82@naver.com, js0ngwn@gmail.com, phia130@naver.com, psy000101@naver.com
jkhong@bu.ac.kr, hjpark@ciotsecurity.com

Crawling Analysis Implementation of Cyber Crime Information in Deep Web Environment

Deok-Hyun Hwang,* So-Young Park,*Ji-Seon Bae,*Song-Ju Jeong,*
Jin-Keun Hong,* Hyun-Joo Park**

*Division. of ICT, Baek-seok University
**CioT, Inc

요 약

본 논문에서는 딥웹 환경에서 사이버 범죄 활동에 대한 정보를 중심으로 분석한다. 분석된 정보는 사이버 수사기관에 범죄 분석을 위한 보조정보로 활용될 수 있도록 지원하는 것과 청소년들의 사이버 범죄에 대한 위중성 및 범법성을 인지시키기 위한 교육을 목적으로 활용될 수 있도록 연구되었다. 따라서 본 논문에서는 크롤링, 파싱, 시각화 3가지 과정을 기반으로 딥웹 환경에서 활동하고 있는 정보를 키워드를 중심으로 수집하고 분석하는 솔루션 환경을 구현하였다. 분석된 정보는 사이버에서 일어나는 많은 범죄활동 가운데 가장 일어나기 쉬운 범죄 유형과 주의 깊게 수사가 이루어져야 할 범죄들을 정리하며, 수사의 방향성을 캐치 할 수 있도록 지원하는 기능을 포함한다.

1. 서론

최근 딥웹에서 소아 성애자 사건 및 N 번방 사건이 발생하며 각종 뉴스의 사회적 화두가 되고 있다. 현재 활용하는 인터넷 웹은 Surface Web 으로 일부분의 정보를 제공하는 영역이다. 이에 반해 딥웹이나 다크웹의 경우 숨겨진 영역으로 수사기관으로 범죄 추적을 회피하기가 수월하여 각종 범죄가 빈번히 일어나는 사이버 공간이다. 딥웹 환경에서는 지금 많은 범죄활동이 일어나고 있는 공간으로(예 무기거래, 청부살인, 생체실험, 아동포르노 등)으로 이에 본 연구에서는 딥웹 환경에서 범죄 추적에 필요한 정보를 주기적으로 모니터링 수집 분석하고자 한다. 더 나아가 수사기관이 범죄 추적에 용이한 최근동향 파악 및 수사 방향설정과 관련된 정보를 수집할 수 있도록 예측을 지원하는 솔루션으로도 활용할 수 있다 [1-3].

2. 관련연구

본 연구에서는 크게 3가지 과정으로 구분된다. 검색 질의 폼, 크롤러 등 관련 구성 및 이를 활용하여 특정 딥웹 페이지를 크롤링하는 크롤링 과정, 크롤링 해온 데이터들을 정제하는 파싱과정, 제반 시각화 기술을 활용한 정제된 데이터들을 바탕으로 분석하여 시각화하는 과정으로 구분된다 [4-5].

(그림 1) 사이버 공간 범죄 활동 뉴스 예

[본 논문은 과학기술정보통신부 정보통신창의인재양성사업의 지원을 통해 수행한 ICT멘토링 프로젝트 결과물입니다]

3. 딥웹 환경 접속 및 정보 수집 분석 방법

본 논문에서는 '코챈' 이라는 딥웹 페이지를 대상으로 활동되고 있는 정보를 분석하였다. 이 딥웹 사이트는 한국인 최대 커뮤니티였던 '아고라'가 폐쇄된 이후(2019)로 가장 큰 규모의 딥웹 사이트이다. 그런데 이 사이트에 접속하는 유저들은 이 곳 '코챈' 사이트에서 여전히 활동 중에 있다(2020).

(그림 2) 딥웹 사이트 정보 수집환경 설정 예

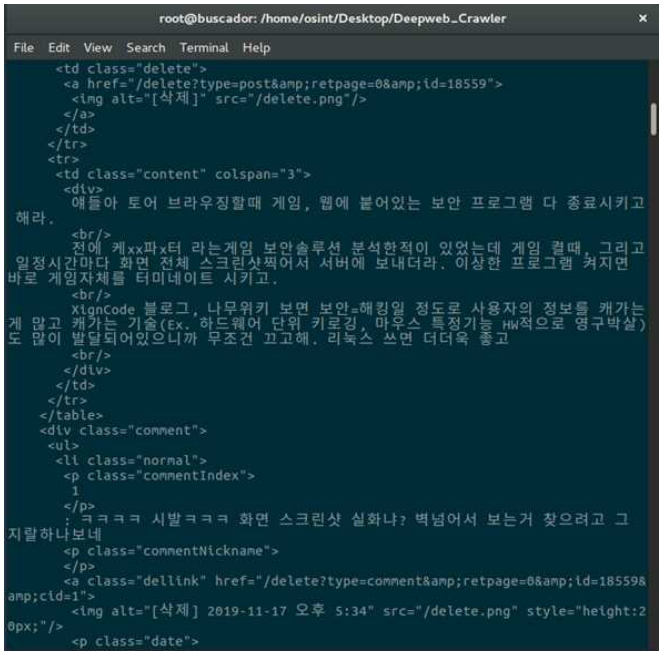
작품 구현 환경은 그림2에서와 같다. 수집 및 분석을 위한 절차는 다음과 같다. 딥웹 페이지 정보수집은 우선 Python의 Requests 모듈을 활용하여 코챈 주소 (http://jqu6my2mlqp4zuui.onion)에 접속하여 관심있는 키워드를 중심으로 크롤링 한다. 이후 크롤링한 키워드 중심의 정보 결과 값을 onion_logs.txt 파일 형식으로 저장 후, 파이썬에서 효율적인 도구인 Parser인 Beautifulsoup4 도구를 활용하여 파싱처리를 수행한다.

크롤링과 파싱을 거친 정제된 데이터들을 기반으로 키워드들의 빈도수를 분석하여 WordCloud 도구를 활용하여 시각화하여 보여준다.

또한 특정 딥웹 url의 파라메터 부분을 반복문으로 실행시키면 원하는 특정페이지들의 데이터들만 크롤링 할 수 있으며, 상기 과정들을 sleep() 함수를 사용해 반복실행을 시키면 주기적인 모니터링이 가능해진다.

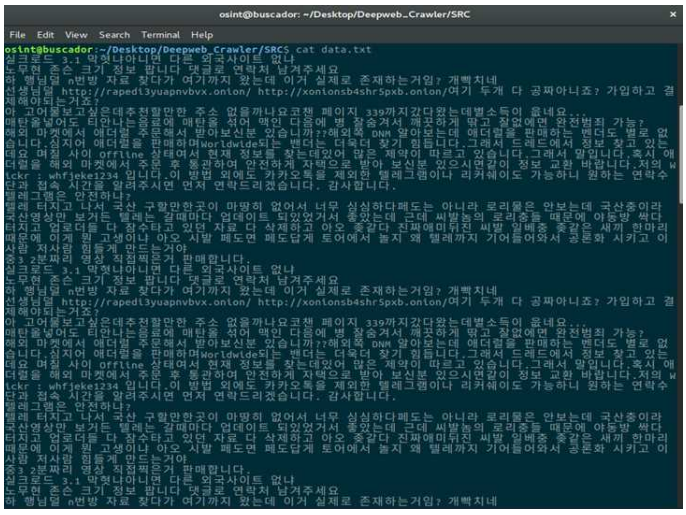
4. 딥웹 환경에서 정보 수집 및 분석 실험 결과

그림 3은 딥웹 환경에서 크롤링하는 장면을 제시한 것이다.



(그림 3) 크롤링 단계 화면

'코챈' 의 웹 페이지에 request를 보내 받아온 html 결과이다. 그런데 불필요한 데이터들(각종html 태그들)이 보이기에 이를 제거하기 위한 파싱과정이 필요하다. 그림4는 수집된 정보를 정제하기 위한 파싱 장면을 제시한 것이다.



(그림 4) 파싱 단계 화면

크롤링 해온 정제되지 않은 데이터들을 Beautifulsoup4 파서를 이용해 파싱을 거친 데이터들이다.

그림5에서와 6은 파싱을 거친 정보에 대해 워드클라우드로 표현한 도식이다.

(그림 5) WordCloud1.

(그림 6) WordCloud2.

쌓여진 파싱된 데이터들의 키워드의 빈도수를 체크해서 WordCloud 도구로 시각화한 결과이다. 빈도수가 높은 키워드들은 키워드의 포인트가 높게 나온다. 제시된 키워드 가운데 가장 큰 키워드가 딥웹 환경에서 가장 많이 나타나는 용어이며, 이 용어와 관련이 있는 상대적으로 높은 연관성을 가진 단어들이 주변에 보다 작게 나타나고 있다.

4. 결론

N 번방 사건을 포함한 관련 사이버 범죄 사건이 지금도 딥웹 공간에서 발생하고 있는 실정이다. 그러나 딥웹에서 유통되는 각종 범죄활동 정보나 관련 정보들이 Surface Web 공간으로 이동에는 일정 시간이 소요된다. 또한 수사기관에서는 딥웹에서 발생하는 각종 정보들을 단시간에 수집하고 분석하는데까지는 한계가 있다. 따라서 본 논문에서는 '제 2의 N번방 사건' 이 될지도 모를 잠재적 위협을 식별 및 분석하여 수사기관에 간접적으로 지원할 수 있는 솔루션 도출을 목표로 하였다. 본 연구결과물로 분석된 데이터들을 토대로 딥웹 사용자들의 현재 관심사를 파악할 수 있으며, 숨겨진 범죄영역을 찾아내기 수월해진다. 수사기관의 딥웹 감시가 명확해질수록 딥웹을 통한 범죄의 감소가 예상된다. 차후 현재 시행하는 많은 단속 중 하나로 자리매김 해 안전한 사이버 환경 제공을 기대할 수 있다. 수사기관에서 이런 방식을 사용해 범죄자를 검거할 수 있음을 교육적으로도 활용할 수 있다. 궁극적으로 수사기관이 범죄 추적에 용이한 최근 동향 파악 및 수사 방향설정과 관련된 정보 제공을 통한 수사 예측 효과를 개선할 수 있을 것이다.

참고문헌

[1] 송지환, 이윤준, 최동훈, "그리드 컴퓨팅 기반 웹 크롤러 시스템 및 그 방법," https://patents.google.com/patent/KR100875636B1/ko

[2] "효율적인 다크웹 모니터링을 위한 탐지 및 색인/검색 기술," https://ictbay.iitp.kr/techdb/ commercialization /getDetailPopView.do;jsessionid= BC627B370B3283C19BB08BA49A4B7D03?NOTI_TECH_ID=75ECM9IZH04VA9E000&PJT_ID=

[3] 문현수, 이영석. "Socks5 프록시 서버를 활용한 토르(Tor)기반 다크 웹 수집 성능 개선," 한국정보과학회 학술발표논문집, pp.912-914, 2019.12.

[4] 조경룡, 조성언, 박장우, "불완전 XML을 위한 파싱 방법," 한국정보통신학회논문지, 12(12), pp. 2153-2158, 2008.12.

[5] 민진우. "뉴럴 전이 기반 한국어 형태소 분석 및 의존 파싱 통합모델," 국내석사학위논문, 전북대학교 일반대학원, 2019. 전라북도.