웹 크롤러 기반 온라인 불법거래 이용자 판단 방법론

김경원*, 신현후*, 정준호*, 손윤식*
*동국대학교 컴퓨터공학과
e-mail : sonbug@dongguk.edu

A Study on the User's Judgment of Illegal Online Transactions Using Web crawler

Kyung-Won Kim*, Hyunhu Shin*, Junho Jeong*, Yunsik Son*
*Dept. of Computer Science and Engineering, Dongguk University, Seoul, Korea

요 약

딥 웹에서는 익명 암호 화폐인 비트코인으로 거래되는 음란물, 위조신분증, 마약과 같은 불법거래가 일어나고 있다. 또한 서피스 웹에서 인터넷과 SNS를 이용한 홍보를 통하여 딥 웹에서만 일어나던 불법거래에 대한 정보가 서피스 웹에 노출되기 시작하였다. 그로인한 불법거래의 규모가 점점 늘어나며 일반인들에게 확산되고 있다. 본 논문에서는 딥 웹의 거래에 사용되는 관리자나 이용자의 SNS ID, 게시글 및 댓글, 홍보글의 텍스트 유형과 URL을 수집하고 그 결과를 서피스 웹에 대조한 유사도를 기반으로 불법거래 이용자를 판단할 수 있는 방법론을 제안한다.

1. 서론

현재 인터넷과 SNS를 이용하여 음란물, 위조신분증, 마약과 같은 불법거래가 늘어나는 추세이다[1]. 딥 웹에서 존재하던 불법거래는 규모가 점차 확장되어 서피스 웹의 SNS를 이용하여 홍보의 수단으로 사용되고 있으며 수사기관의 추적을 피하기 위해 카카오톡, 텔레그램, 트위터를 이용해 사이트를 공유하거나 거래를 유도한다. 또한 실질적 거래는 차명계좌, 비트코인을 통하여 하는 등의 거래 방법 또한 다양해지고 있으며 이용자는 지속적으로 증가하고 있다[2]. 관련 수사기관은 불법거래에 대한 관리의 필요성을 인식하고 함정수사 및 비트코인 거래내역을 추적하는 방식을 이용하여 수사 중이다. 그러나 범죄규모에 비해 인력과 기술력이 미흡한 실정으로 어려움이 존재한다[3].

본 논문에서는 딥 웹 크롤러를 이용하여 불법거래에 사용되는 관리자와 이용자에 대한 SNS ID, 게시글 및 댓글, 홍보글의 텍스트 유형, URL에서 데이터를 수집하고 서피스 웹과 대조한 유사도 기반의 불법거래 이용자를 판단할 수 있는 방법론을 제안한다.

이 성과는 2018년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행되는 연구임 (No.2018R1A5A7023490).

2. 관련연구

일반 검색엔진인 네이버, 구글 등을 이용하여 접속하는 웹 사이트는 서피스 웹이라 하고, 반대로 일반적인 검색엔진에서 접속할 수 없는 사이트를 딥 웹이라고 한다. 서피스 웹의 규모는 웹 전체 온라인 컨텐츠의 4%이고 나머지는 딥 웹이라는 조사 결과가 있다[3].

딥 웹은 접속 허가가 필요한 네트워크나 특정 소프트웨어를 이용하여 IP주소를 몇 차례 우회하는 오버레이 네트워크이다. 따라서 프록시 우회접속 도구를 사용하지 않으면 접속하기 힘든 구조이다. 그러므로 네트워크상에서 딥 웹의 이용자 판단이 매우 힘든 상황이다[4].

Tor는 딥 웹의 네트워크 우회와 익명화를 위해 사용되는 대표적인 프록시 서버 우회접속 도구이다[5]. Tor의 프록시 서버 우회는 인터넷에서 사용자간의 익명성을 보호해주기 위한 여러 자원봉사자들이 노드가 되어 운영하는 서버이며 이러한 노드를 3회에 거쳐 접속하게 되어있다. 주소에 .onion이라는 고유의 도메인을 가진 딥 웹은 Tor를 이용하지 않으면 접근할 수 없는 체계를 가지고 있다.

웹 크롤러는 방대한 웹에서 자동화된 방법으로 특정 웹 문서를 탐색하여 수집하는 기술을 말한다[6]. 검색 엔진과 같은 방식으로 여러 사이트에 방문하여 페이지의 HTML, XML 등 웹의 소스코드를 분석하고 원하는 형태의 정보를 수집하는 데에 사용된다. 이러한 작업을 웹 크롤링 또는 스파이더링이라 지칭하기도 한다. 웹 크롤링은 정해진 규칙에 따라 크롤링을 하므로 불규칙적인 딥 웹의 정확한

크롤링을 위해선 많은 연구가 필요하다.

3. 웹 크롤러 기반 온라인 불법거래 이용자 판단 방법론

아래 그림 1은 딥 웹에서 다른 딥 웹의 주소 공유를 위한 곳으로 일반인들도 Tor에서 검색만으로 들어올 수 있는 한국 위키이다. 딥 웹은 다양한 딥 웹의 주소를 제공해주는 위키가 존재한다. 주로 위키의 항목들은 블랙마켓, 딥 웹 커뮤니티, 대리 해킹, 음란물, 마약 등 대부분 불법거래에 관한 항목이다. 서피스 웹과 딥 웹의 경계가 없어져 가는 상황에서 본 논문에서 제안하는 불법거래 이용자를 판단할 수 있는 기술의 필요성이 증대된다.

그림 1. 딥웹 마약판매 사이트 모음

딥 웹 크롤러는 크롤링을 가능하게 하는 환경설정이 필요하다. 파이썬의 라이브러리인 SOCKS를 사용할 수 있다. SOCKS는 TCP 프록시 기능을 지원해주며 프록시 서버에 Tor IP와 port를 적용해서 네트워크를 설정하게 되면 딥 웹의 데이터를 검색 및 수집할 수 있는 크롤링을 설정할 수 있게 된다[7].

먼저, 딥 웹의 경우 대표적 카테고리인 블랙마켓, 아동, 딥 웹 커뮤니티, 마약, 대마초 등 범죄관련 중심으로 크롤링을 한다. 게시글 및 댓글에는 비트코인 지갑의 주소와 SNS ID 혹은 다른 웹으로 유도하는 데이터를 수집할 수 있으며, 홍보글의 텍스트 유형과 URL 데이터를 수집하면서 영역을 넓혀 크롤링을 한다.

서피스 웹은 딥 웹 크롤러의 결과에서 자주 사용되는 키워드와 관련된 모든 데이터를 수집한다. 최종적으로 두 결과를 대조하여 이용자나 홍보글에 대한 유사도를 분석할 수 있게 된다. 일치하는 이용자의 ID나 홍보글, URL이 존재한다면 일반인들이 딥 웹에 유입되는 경로와 딥 웹과 서피스 웹 양쪽에서 불법거래를 하고 있는 이용자를 판별할 수 있다. 이러한 정보를 제공함으로써 관련 수사기관에게 도움을 줄 수 있을 것으로 판단된다.

본 논문에서 제안하는 방법론에서 크롤링한 텍스트는 크게 두 가지 종류로 구분할 수 있다. 이용자 ID와 홍보글 텍스트이다. 이용자 ID같은 경우 최대 ID 길이가 제한되어있으며 제한된 길이를 넘었을 경우 홍보글로 인식하여 분류한다.

아래 그림 2는 제안하는 방법론의 불법거래 이용자 판단을 위한 웹 크롤러 체계를 그림으로 나타내었다.

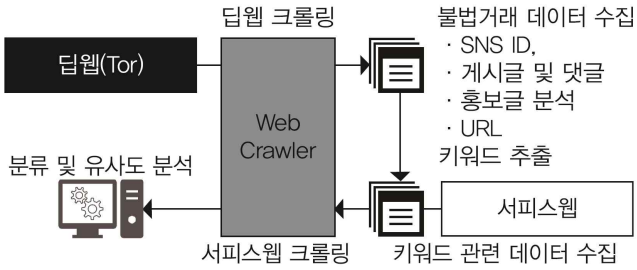

그림 2. 불법거래 이용자 데이터 웹 크롤러 체계도

4. 판단 방법

ID와 홍보글의 유사도 산출은 편집거리 알고리즘(Levenshtein Distance, Edit Distance)의 결괏값을 이용한다. 아래 표1은 편집거리 알고리즘의 의사코드 형식으로 작성했으며, 표2로 예시를 들어 설명하려 한다.

표 1. 편집거리 알고리즘 의사코드

```
1: A = string1, B = string2
2: ARR = [][]
3: MAXLEN = MAX STRING LENGTH(A, B)
4:   for(i=1; i<=A.LENGTH; i++) ARR[i][0] = A[i-1]
5:   for(j=1; j<=B.LENGTH; j++) ARR[0][j] = B[j-1]
6:   for(j=1; B.LENGTH; j++){
7:     for((i=1; A.LENGTH; i++){
8:     if(A[i-1] == B[j-1])
9:       ARR[i][j] = ARR[i-1][j-1];
10:    else
11:      ARR[i][j] = min(ARR[i-1][j-1]+1, min(
         ARR[i][j-1]+1, ARR[i-1][j]+1))}}
```

표 2. 편집거리 예시

	m	u	i	t	l
m	0	1	2	3	4
u	1	0	1	2	3
l	2	1	1	2	3
t	3	2	2	1	2
i	4	3	3	2	2

편집거리 알고리즘은 두 문자열이 같아지기 위해 최소 몇 번의 변경, 추가, 삭제를 필요로 하는지에 대한 최소 횟수를 산출하는 알고리즘이다.

표 2는 과정을 표현하였다. 첫 행과 첫 열에 비교할 문자열을 배치한다. 각 칸마다 있는 코스트 값은 같은 문자가 없거나, 수정해야하거나, 추가해야할 때 1씩 늘어난다. 위 배열의 마지막인 배열[5][5]는 편집거리를 계산하고 나온 값이 최소편집거리가 된다.

아래 표3은 유사도를 계산하는 알고리즘 의사코드다.

표 3. 최소편집거리를 이용한 유사도 산출 의사코드

```
1 : A = Deep Web String
2 : D,E,F = Surface Web String
3 : M = Minimum Distance(A, D), (A, E), (A, F)
4 : MAXLEN = MAX STRING LENGTH(A, B)
5 : SYM[] = ((MAXLEN-M) / MAXLEN) * 100
6 : for(i=0; Number of Surface Web strings; i++)
7:       ID[i] = M
8 : Repeat. Compare all Surface Web strings.
9 : min(ID.index)
```

유사도 산출은 편집거리 알고리즘의 결괏값인 최소편집거리를 이용한다. 최소편집거리는 두 문자열이 같아질 때까지의 추가, 변경, 삭제한 횟수를 나타낸다. 즉, 값이 낮을수록 두 문자열이 유사하다는 것을 뜻한다. 표 2를 예를 들자면, 최대 길이가 5인 두 문자열에서 최소편집거리는 2이다. 표 3의 5라인을 참고하면 ((5-2) / 5) * 100이 되며 결괏값으로는 60이라는 유사도가 나온다. 딥 웹의 문자열에 다수의 서피스 웹 문자열에 대한 최소편집거리 값들을 배열에 저장하고 최소값의 인덱스를 선정한다.

딥 웹 사이트의 홍보글은 판매자마다 일정한 양식을 유지하나, 딥 웹ID는 완전히 일치하게 만들지 않는다. 이는 딥 웹 관리자가 회원가입 시 평소 습관대로 ID를 기입하지 말 것을 강조하기 때문이다. 그러나 불법거래 이용자들은 SNS ID와 텔레그램 혹은 카카오톡 ID 자체를 비슷한 유형으로 만드는 경향이 있다. ID의 앞부분은 같으나 넘버링을 추가하거나 변경하는 등의 유형을 많이 이용하므로 편집거리 알고리즘을 이용한 대상 텍스트간의 유사도를 산출할 필요성이 있다. 딥 웹의 ID가 다르더라도 딥 웹과 서피스 웹에서 자주 이용하는 거래수단의 ID는 동일할 것이다. 따라서 두 문자열에 대해 편집거리 알고리즘을 이용하여 가장 높은 유사도를 가진 텍스트와 연관된 정보를 이용하여 판단할 수 있다.

5. 결론 및 향후연구

특정 소프트웨어만 설치하면 누구나 접근이 가능한 딥 웹을 일반인들이 관심을 가지기 시작하면서 불법적인 자료와 거래들이 점점 늘어나고 있으며 이를 억제하기 위한 대응방안이 부족하다. 그러므로 자동으로 데이터를 수집하고 판단 할 수 있는 웹 크롤러의 필요성이 증대된다.

본 논문에서는 온라인상에서 일어나는 불법거래의 이용자를 판단하기 위해 딥 웹에서 Tor 네트워크를 이용한 크롤링 결과를 토대로 서피스 웹과 대조한 유사도를 기반으로 불법거래 이용자를 판단할 수 있는 방법론을 제안했다.

향후 연구로는 제안한 크롤러의 기능적인 부분의 구현과 딥 웹의 원활한 크롤링을 위해서 다양한 회원제 가입 및 비트코인 지갑의 키를 이용한 회원인증 등의 여러 유형을 극복하는 방법에 대해 연구가 필요하다. 또한 불법거래 이용자 판단 기능의 향상방안에 대해 연구하고자 한다.

참고문헌

[1] 박호정, “인터넷과 sns를 이용한 마약거래 대응방안에 관한 연구”, 융합보안논문지, pp. 93-102. 2018.

[2] Bin He, Mitesh Patel, Zhen Zhang, Kevin Chen-Chuan Chang, "Accessing the Deep Web : A Survey", University of Illinois at Urbana-Champaign, pp. 1-8, 2007.

[3] 박웅신, 이경렬, “다크넷 범죄현상과 형사법적 대응방안”, 대검찰청, pp. 223-224, 2018.

[4] 남기천, 이윤호, “딥웹상 범죄에 대한 함정수사 도입에 관한 연구”, 한국공안행정학회, pp. 84-90, 2016

[5] 이경빈, “토르 히든서비스 서버 분석 사례 및 추적기법에 관한 연구”, 경찰청, pp. 38-45, 2018.

[6] 김광영, 이원구, 이민호, 윤화묵, 신성호, “웹 자원 아카이빙을 위한 웹 크롤러 연구 개발”, 한국콘텐츠학회논문지, 제11권, 제9호, pp. 10-11, 2011.

[7] ProxyBroker, “https://github.com/constverum/ProxyBroker”,github.