Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.26, NO.2, Apr. 2016 ISSN 2288-2715(Online)
http://dx.doi.org/10.13089/JKIISC.2016.26.2.397

단어 조합 검색을 이용한 불법·유해정보 탐지 기법

한 병 우,† 윤 지 원‡
고려대학교

Illegal and Harmful Information Detection Technique Using Combination of Search Words

Byeong Woo Han,† Ji Won Yoon‡
Korea University

요 약

최근 국내에서 불법·유해정보의 양은 꾸준히 증가하고 있으며, 중소기업, 공공기관 등의 게시판에 불법·유해정보 글들이 많이 게시되고 있다. 불법·유해정보를 통해 범죄로 이어질 가능성이 크기 때문에 이를 탐지하는 시스템이 필요하다. 현재 국내의 불법·유해정보 탐지는 인력에 의해 수동적으로 진행되고 있다. 본 논문에서는 공개출처정보(OSINT)를 통해 불법·유해정보 중 마약 판매 게시글의 URL 탐지를 자동화하는 연구를 진행하였다. 이 시스템은 마약 판매 게시글의 단어를 분석하고, 해당 단어로 검색어 사전을 만들었다. 검색어 사전 기반으로 검색되는 마약 판매 의심 URL을 구글 검색엔진을 활용하여 자동으로 수집하였다. 수집 URL을 도메인별로 분류하였으며, 도메인을 수집 URL 개수별로 도식화하여 실제 불법·유해정보를 찾아내었다. 이 자동화 탐지 시스템을 활용하면 모니터 요원의 수동적인 탐지업무로 인한 시간과 노력의 소비 문제를 해결할 것으로 기대된다.

ABSTRACT

Illegal and harmful contents on the Internet has been an issue and been increased in Korea. They are often posted on the billboard and website of small enterprise and government office. Those illegal and harmful contents can relate to crime and suspicious activity, so, we need a detection system. However, to date the detection itself has been conducted manually by a person. In this paper, we develop an automated URL detection scheme for detecting a drug trafficking by using Google. This system works by analyzing the frequently used keywords in a drug trafficking and generate a keyword dictionary to store words for future search. The suspected drug trafficking URL are automatically collected based on the keyword dictionary by using Google search engine. The suspicious URL can be detected by classifying and numbering each domain from the collection of the suspected URL. This proposed automated URL detection can be an effective solution for detecting a drug trafficking, also reducing time and effort consumed by human-based URL detection.

Keywords: OSINT, Inteligence, illegal drug, Google, Security, Criminal detection, Automatic detection

I. 서 론

최근 파리테러(2015)이후로 사이버 인텔리전스(Cyber Intelligence)에 대한 관심이 증가되고 있다. 파리테러이후 핵티비즘(Hacktivism) 단체인 어나니머스(Anonymous)는 ISIS관련 사이트를 공격하여 관련 정보를 공개하였다[1]. 또한, 미국의 경우 다크 웹(Dark Web)상에 있는 테러 정보를 이용하여 테러리스트 조직의 구조를 파악하였다. 이에 활용된 사이버 인텔리전스의 핵심은 인터넷상에 존재

Received(02. 02. 2016), Modified(03. 17. 2016), Accepted(03. 25. 2016)
† 주저자, sosnsos@korea.ac.kr
‡ 교신저자, jwyoon@korea.ac.kr(Corresponding author)

하는 공개출처정보(OSINT: Open Source Intelligence)를 이용한 활동이다.

공개출처정보는 공공에서 접근이 가능한 신문, 연구논문, 인터넷 등과 같은 공개 정보(Open Source)를 통하여 정보를 수집하는 활동(Intelligence)을 의미한다[2]. 사이버 상 공개출처정보의 주요 활동영역으로는 인터넷을 이용한 배경조사, 범죄수사, 정보수집이라는 영역으로 나눌 수 있다[3].

이러한 공개출처정보의 활동은 세계 각국에서 다양한 방면으로 활용되고 있다. 대표적으로 미국의 프리즘(PRISM)을 예로 들 수 있다. 미 국가 안보국(NSA)을 중심으로 국가 보안 전자 감시 체계 중 하나인 프리즘을 이용하여 50회 이상의 잠재적인 테러 공격을 방지하였다[4]. 최근에는 프리즘에 트위터(Twitter)와 페이스북(Facebook)등과 같은 사회관계망서비스(SNS)에 대한 정보를 수집하기 시작하였다. 영국에서는 영국 경시청을 중심으로 IRA(Irish Republican Army)와 같은 국가적 극단주의자들을 선정하여 지속적으로 감시할 수 있는 팀을 조직하여 공개출처정보를 활용하고 있다[5].

현재 국내의 경우, 공개출처정보를 불법·유해정보 탐지에 활용하고 있다. 국내에서 활용되는 불법·유해정보 탐지 시스템중 대표적인 시스템 2가지는 방송통신위원회와 경찰기관의 모니터링 시스템이다. 방송통신위원회의 불법 모니터 탐지 시스템은 자체 모니터 요원을 활용하여 불법·유해정보를 탐지하고 있다. 경찰기관에서는 자체 모니터링 시스템인 "누리캅스"를 활용하여, 불법·유해정보를 탐지하고 있다. 경찰인력 이외에도 민간요원을 선발하여 모니터링을 운영하여 현재 성과를 내고 있다.

하지만 증가하는 불법·유해정보에 비해 모니터 요원이 많이 부족한 실정이며, 대표적으로 운영되고 있는 2가지 탐지 시스템도 대부분 수동적으로 활용되고 있어 인터넷상의 모든 불법·유해정보를 효율적으로 탐지하는 것에는 한계가 존재한다[6]. 예를 들어 소규모 웹사이트들이 점점 증가하고 있지만 자체 필터링 관련 기술 보유와 관리 소홀 등으로 인해 불법·유해정보가 무분별하게 업데이트 되고 있어 이를 수동적으로 탐지하기에는 인력적 자원의 한계가 있다.

이에 본 논문에서는 검색엔진을 이용하여 찾을 수 있는 공개출처정보를 이용하여 효율적으로 수집할 수 있는 자동화시스템을 연구하였고, 이를 현재 국내 인터넷상에 있는 불법·유해정보 중 마약 판매에 초점을 맞추어 실험을 진행하였다.

본 논문은 2장에서는 공개출처정보와 관련된 연구에 대해 설명하였고, 3장에서는 본 논문에서 불법·유해정보 자동화 탐지 시스템에 대하여 설명하였다. 4장 탐지 자동화 결과에 대하여 설명하였다. 5장에서는 결론 및 향후 연구 방향에 대하여 설명하였다.

II. 관련 연구

2.1 공개출처정보(OSINT)

2001년 9.11테러 이후, 테러리스트 및 범죄 탐지를 위해 공개출처정보와 이와 관련된 연구가 활발히 진행되어 왔다. 공개출처정보와 관련되어 지금까지 진행되어온 연구 중 대표적인 미디어 분석과 관계망 분석 연구에 대해 설명하고자 한다.

첫째, 미디어 분석을 통한 공개출처정보 연구는 2008년 Clive Best 연구진의 연구에서는 EMM(Europe Media Monitor)의 인터넷상 공개 출처 정보인 뉴스를 수집·분석하여 이슈를 실시간으로 탐지 가능한 경고 시스템에 대하여 연구하였다. 이 경고 시스템은 목표로 설정된 이슈가 발견되면 자동적으로 관련된 주제 및 이와 관련된 인적사항 등의 정보를 제공한다. 또한, 이 시스템은 인터넷 지도를 활용하여 사건에 대한 지리적 위치를 확인 할 수 있게 하였다. 2011년 Clive Best 연구진은 2008년의 기존 연구에 사회관계망 미디어도 분석을 추가하여 진화된 시스템을 만들었다. 이 연구에서는 자동분석시스템의 정보 진위여부에 대한 검증을 중점으로 연구하였다. 하지만, 정보의 진위여부는 100% 자동화 할 수 없기 때문에 사람의 판단이 필요하다고 한다[7][8]. 2008년에는 F.Neri 연구진이 선택한 목표의 인터넷 상에 존재하는 공개정보를 수집, 분석 그리고 분류하는 연구를 진행하였다. 이 시스템의 가장 큰 특징은 목표를 설정하여 수집된 정보에서 파생된 또 다른 관련 정보를 확장하여 수집 및 분석 한다는 점에 있다[9].

두 번째로 최근 사용 빈도가 높은 사회관계망 서비스 분석을 통한 공개출처정보의 연구가 있다. 사회관계망 서비스분석은 대부분 부족한 데이터를 활용하여 분석을 한다. 2011년 Christopher J.Rhodes 연구진은 단편적이고 제한적인 데이터를 가지고 있는 사회관계망서비스의 정보를 수집하였다. 이 연구에서는 제한된 데이터를 효율적으로 분석하기

위해 Data-hungry 기술을 사용하였다. 핵심적인 역할을 하지만 드러나지 않는 중요한 사용자들(Key players)을 찾아내는 연구를 진행하였다[10]. 2012년에는 Paul A. Watters 연구진이 사회 관계망 서비스(SNS)상에서 불법 마약 거래를 하는 개인과 범죄단체 등을 탐지하는 부분적 자동화 시스템을 만들었다. 이 연구에서는 공개출처정보의 분석을 통해 호주 내의 마약 판매 분포를 찾아내었다[11].

현재까지 진행되어온 공개출처정보 관련 연구들은 정보분석 및 평가에 중점을 두고 이를 효율적으로 탐지하는 방법 초점을 두었다. 하지만, 본 논문에서는 이들 정보를 자동적으로 탐지 후 수집하는 방법에 대하여 제안하고자 한다.

2.2 TF-IDF

TF-IDF(Term Frequency - Inverse Document Frequency)는 정보 검색과 텍스트 마이닝에서 활용하는 가중치이다. 문서군이 있을 때 어떤 단어가 문서 내에서 중요도를 나타내는 통계적 수치이다. 문서의 핵심 단어 추출이나, 문서들의 유사정도를 계산하는 등의 용도로 사용할 수 있다[14].

단어 빈도(TF, term frequency)는 문서 내에 특정 단어의 빈도를 나타내는 값으로, 이 값이 높을수록 단어가 문서에서 중요하다 볼 수 있다. 하지만 단어가 문서군 내에서 자주 사용되면 흔하게 사용되는 것을 의미한다. 이것을 문서 빈도(DF, document frequency)라고 하며, 특정 단어가 나타난 문서의 수를 나타낸다. 문서 빈도 값이 클수록 문서 내에서 해당 단어가 중요하지 않다는 것을 나타낸다. 문서 빈도 계산 시 값의 편차가 매우 크게 되므로 이 차를 줄이기 위하여 로그(log)값을 취한다. 이 수치의 역수를 역문서 빈도(IDF, inverse document frequency)라고 한다. TF-IDF는 두 값을 곱하여 산출하는 방식으로 이 값을 활용하면 흔하게 사용되는 단어를 걸러낼 수 있는 효과를 얻을 수 있다.

$$tf(t,d) = f(t,d) \quad (1)$$

$$idf(t,D) = \log \frac{|D|}{|\{d \in D : t \in d\}|} \quad (2)$$

$$tfidf(t,d,D) = tf(t,d) \times idf(t,D) \quad (3)$$

해당 식의 문자에서 t는 특정 단어, d는 문서, $|D|$는 전체문서의 수를 나타내며, $\{d \in D : t \in d\}$는 단어 t가 포함된 문서의 수를 나타낸다. 식(1)은 문서 내에 단어의 빈도를 계산하는 식이며, 식(2)은 전체의 문서 군에서 해당 단어가 포함된 문서의 빈도를 계산한 후 역수를 취한 값이다. 식(3)은 식(1)과 식(2)을 곱하여 나온 값으로 특정 단어가 특정 문서에서 얼마나 중요도를 계산할 수 있다. 이 계산 값을 이용하여 문서 내에 단어들의 중요도를 정렬 할 수 있다.

III. 불법·유해정보 탐지 자동화

본 연구에서 공개출처정보의 자동화에 중점을 둔 시스템으로 총 6개의 단계로 구성되어 있다. 본 시스템은 구글 검색엔진을 기반으로 ① 게시글 샘플 추출, ② 게시글 샘플 분석 및 단어 사전 생성, ③ 검색어 사전 생성, ④ 검색 및 결과 정보 저장, ⑤ 분석, ⑥ 도식화 6단계로 나누어지며, 이는 Fig.1.과 같다.

불법·유해정보 자동화 탐지 시스템을 실제 마약 판매 사이트를 탐지하는 것에 목표를 두고 실험을 진행하였다. 본 연구에서는 실험을 진행하기 전에 마약

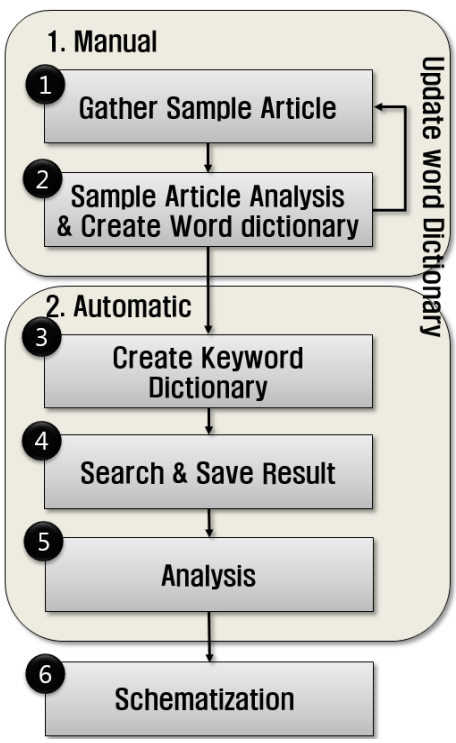

Fig. 1. Detection system process

판매 글에 대한 불법·유해정보는 마약을 구하려는 불특정 다수의 사람들을 목적으로 작성한다는 것에 가정을 두었다.

프로그램 개발에 사용한 언어는 Java 1.7이며, 자동화 테스트 프레임워크인 Selenium 2.48.2, 브라우저는 Firefox 43.0을 활용하였다.

3.1 게시글 샘플 수집

검색어 샘플 추출 단계에서는 마약 관련 단어를 검색하였을 때 검색 상단에 노출되는 웹사이트를 중점으로 게시글을 추출하였다. 게시글 샘플은 2014년과 2015년에 작성된 글을 기준으로 30개의 게시글을 수집하였으며 검색 시 날짜가 없는 글들은 수집하지 않았다. 국내 대형 포털에서 운영하는 게시판 및 블로그 등은 내부 필터링과 모니터링이 이루어져 마약 판매 게시글의 검색이 어렵지만, 이외의 웹사이트들은 기술력과 관리 부족으로 인해 판매 게시글을 수집이 가능하였다[11]. 특히, 마약 판매 게시글은 웹사이트 관리가 제대로 이루어지지 않거나 미가입 작성 게시판을 중심으로 작성되어져 있는 경우가 많다.

3.2 게시글 샘플 분석 및 단어 사전 생성

본 실험에서는 수집된 30개의 마약 게시글에서 검색엔진에 사용될 단어를 만들기 위해 출현 빈도가 높은 단어를 추출하여 분석하였다. 본 실험에서 게시글의 HTML 태그(Tag)를 제거하기 위해 오픈소스 HTML 파서(Parser) 라이브러리인 Jsoup을 활용하였으며, 단어 분석은 서울대학교에서 개발한 꼬꼬마 한글 형태소 분석기의 기능 중 색인어 추출을 활용하였다[12][13].

Fig. 2와 같은 방식을 통해 단어 사전을 생성한다. 게시글 본문을 Jsoup의 HTML 태그 제거 기능을 활용하여 게시글 본문에 존재하는 HTML 태그를 제거하였다. 형태소 분석기를 이용하여 추출한 색인어들은 TF-IDF 알고리즘을 이용하여 연산하였으며, 각 게시글 당 TF-IDF 값이 높은 10개의 단어들을 추출하였다[14]. 각 문서 당 추출된 단어들 중 '2010', '허브팝', '팝' 등과 같은 의미가 없는 단어들이 나와 분석기의 성능은 완벽하지 않아 분석자의 판단에 의해 수동으로 제거하였다. 또한, 마약 판매의 경우 일반어를 은어로 사용한 경우가 많기 때문에 분석자의 판단에 의해 단어들을 선택하였다. 가령

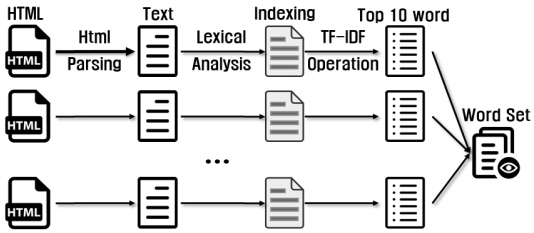

Fig. 2. Word set creation process

'작대기'라는 단어는 마약 은어에서는 '주사기'를 뜻하는 이중적인 의미를 가진다. 사전적으로는 '긴 막대기'를 뜻하지만 마약사범들에게는 '마약을 주입하는 주사기'를 뜻한다. 그래서 단어 사전 생성 시 필요 단어가 누락되지 않게 사람의 판단이 필요하다.

이 과정을 통하여 추출한 단어 는 총 20개 이며 "허브, 크리스탈, 작대기, 아이스, 물뽕, 도리, 여성흥분제, 얼음, 필로폰, 술, 히로뽕, 수면제, 순도, 마리화나, 최상급, 대마초, 쿠쉬, 청산가리, 시안화칼륨, 엑스터시" 단어들을 추출하여 단어사전을 만들었다.

3.3 검색어 사전 생성

추출된 단어사전을 이용하여 단어를 3개씩 조합(Combination)한 검색어 사전을 생성한다. 3개씩 단어를 조합한 이유는 앞서 3.2에서 설명하였 듯이 마약 판매에 사용되는 단어들은 일반단어를 은어로 활용하기 때문이다. 여러 단어를 조합할 경우, 하나의 단어를 이용하였을 때보다 상대적으로 정확한 검색결과를 얻을 수 있다[6]. Table 1.은 단어사전 중 '허브, 크리스탈, 작대기'를 구글에 검색한 웹페이지의 검색결과이다. 한 개의 단어와 단어 3개를 조합하여 검색하였을 때 마약판매 게시글이 얼마나 검색되는지 개수를 세었다. Table 1.에서 나온 검색결과를 보면 단어를 하나만 검색하였을 때보다는 조합하였을 때가 더 많은 마약 판매 게시글이 검출되는 것을 확인 할 수 있다. 단어 한 개를 검색 시 '작대기'을 제외하고는 모두 일반 게시글이 검색되었다. 36개의 마약 판매 게시글이 나온 '작대기'를 검색하였을 때도 조합한 검색어의 마약판매 게시글보다 낮은 검색결과를 확인 할 수 있다.

위에서 추출된 단어들을 이용하여 단어를 3개씩 조합(Combination)하여 {허브, 크리스탈, 작대기}, {허브, 크리스탈, 아이스}, {허브, 크리스탈, 물

Table 1. Combination and Non-Combination search result

Keyword	Drug	Normal
허브	0	100
크리스탈	0	100
작대기	36	64
허브 크리스탈	31	69
허브 작대기	99	1
크리스탈 작대기	96	4
허브 크리스탈 작대기	100	0

뽕)... 등의 총 1140(20C3)개의 검색어를 가지는 검색어 사전을 생성하였다.

3.4 검색 및 결과 저장

생성된 검색어 사전을 활용하여 검색엔진 중 하나인 구글에서 정보 검색을 수행하였다. 구글은 자동화 판단을 위해 캡차(Captcha)가 실행되어 검색결과의 수집을 막고 있다. 캡차를 극복하기 위해 구글 검색 대행 사이트인 Disconnect.me를 활용하였다. 검색 결과 수집을 자동화하기 위하여 소프트웨어 테스트 프레임워크인 Selenium을 활용하였다[15].

Fig.3.과 같이 Java Thread를 이용하여 동시에 3개의 브라우저를 실행하였다. 검색어 사전을 활용하여 각 브라우저에서 마약관련 정보 검색을 수행하였다. 검색어 사전을 이용하여 검색된 결과를 최대 10Page까지의 검색결과를 수집(Crawling)하였다. 검색결과 중 제목, URL, 요약을 파싱(Parsing)하여 데이터베이스에 저장하였다.

한번 수집을 수행하는 데 약 2시간 정도 수집을 하게 되며, 약 7천에서 4만건 사이의 검색 결과를 수집한다. 수집 시 검색결과가 일정하지 않은 이유는 구글 검색 대행 사이트인 Disconnect.me가 수집 실행마다 일정한 검색결과를 돌려주지 않기 때문에 검색 결과가 일정치 않다. 향후 검색결과를 온전히 돌려줄 수 있는 시스템 보완이 필요하다.

3.5 분석

데이터베이스에 저장된 마약관련 데이터가 게시된 URL은 SQL쿼리를 이용하여 중복 없이 컬럼(Column)을 추출하였다. 특정한 규칙을 가진 문자열의 집합을 표현하는 데 사용하는 형식언어인 정규표현식을 이용하여 URL에서 도메인을 추출하였으

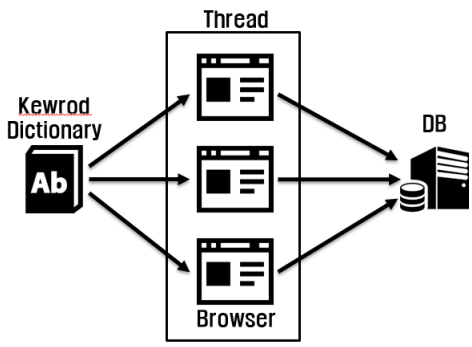

Fig. 3. Search collection process

다. Fig.4.와 같이 URL과 도메인은 각자의 테이블에 저장을 하였으며, URL을 검색할 때 사용하였던 검색어는 키워드 테이블에 저장을 하였다. 도메인 분류 과정에서 기존에 저장되어 있지 않으면 도메인, URL, 키워드 테이블은 새로 컬럼을 생성하여 저장하고, 도메인테이블에는 수집된 URL 개수를 저장한다. 추후 새로운 키워드로 URL이 검출되면 키워드 테이블도 업데이트 된다. 정규표현식으로 도메인을 분류 시 www.daum.net, daum.net은 같은 웹사이트를 보여주는 도메인이지만 이 시스템에서는 서로 다른 도메인으로 규정을 하였다. 도메인, URL, 키워드 테이블은 도식화를 위해 서로 연관되어 있다.

3.6 도식화

도식화는 앞서 전처리 과정을 통하여 도메인테이블에 저장된 수집된 URL 개수가 많은 순으로 오름차순으로 정렬되어 Web을 활용하여 도식화 하였다. 도메인 목록 화면은 한 페이지에 총 20개의 도메인들을 보여주며, 도메인 아이디, 도메인 명, 시스템에 도메인 등록일자, 수집된 URL 개수를 확인할 수 있

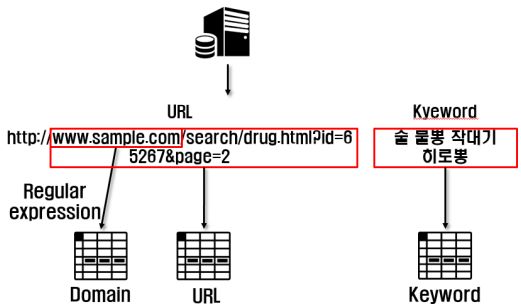

Fig. 4. Analysis

다. 또한, 수집된 URL 수을 기준으로 각 도메인을 정렬하여 출력한다. 도메인을 클릭하였을 때 수집된 URL 목록 화면으로 이동한다. 수집 URL 목록화면에서는 제목, 키워드, 키워드 개수를 확인 할 수 있다. 제목을 클릭하면 마약 판매 의심 게시글로 이동되며, URL 수집 시 사용하였던 키워드와 키워드 개수를 확인 할 수 있다.

IV. 불법·유해정보 탐지 자동화 결과

앞서 Fig.5.는 수집된 URL을 도메인별로 분류하고 수집 URL을 기준으로 상위 10개의 사이트를 도식화한 화면이다. 각 도메인을 클릭하면 도메인에 해당하는 Fig.6.과 같은 도메인에 해당하는 수집 URL 목록을 확인 할 수 있다. 상위 10개 중 blog.naver.com, blog.daum.net 등은 국내 대형 IT 업체가 운영하는 도메인이다. 국내 IT업체 도메인의 수집된 URL 링크을 전수 조사하였을 때 마약을 설명하거나 일반 게시글로 마약 판매글이 존재하지 않는 것으로 확인되었다. 하지만 나머지 상위 10개의 도메인은 마약 판매 의심 게시글을 다수 확인 할 수 있다. Fig.6.은 'Youtube' 도메인을 클릭하였을 때 나온 마약 판매 의심 URL 목록이며, Fig.7.처럼 URL 링크를 클릭하면 일반인들도 게시글에 접근 가능한 것을 확인할 수 있다. 시스템을 통해 대부분의 마약 판매글들은 웹사이트 관리가 제대로 이루어지지 않거나 미가입 작성 게시판을 중심으로 작성되어져 있는 경우가 많은 것으로 판단되었다.

DomainId	도메인	등록일자	URL 수
6670	blog.naver.com	2015.12.09	221
5781	www.jisanchem.com	2015.12.09	176
5993	www.qldkorean.net	2015.12.09	144
5581	www.youtube.com	2015.12.09	132
5892	dias.pknu.ac.kr	2015.12.09	107
5846	www.overwork.or.kr	2015.12.09	106
5628	blog.daum.net	2015.12.09	104
5588	arch.mutodesign.co.kr	2015.12.09	103
5777	m.blog.daum.net	2015.12.09	88
5582	sk9.kr	2015.12.09	64

Fig. 5. Domain list

NO	TITLE	Keyword	키워드개수
15879	얼음 팝니다 아이스 팝니다 크리스탈 팝니다 작대기 팝니다 술 ...	뽕,술,아이스,떨,크리스탈,얼음,작대기,빙두,허브,필로폰,물뽕히로뽕히로뽕	11
15944	하율이의 도리도리 댄스 - YouTube	뽕,술,도리도리	3
16305	도리도리 대마초판매 카톡 angel7000 헤시시 얼음 물뽕 ghb ...	술,떨,ghb,얼음,허브,물뽕,아이스,빙두,대마초,크리스탈,작대기,도리도리,헤시시,필로폰	14
16461	카톡jayk1 물뽕 수면제GHB 아이스 작대기 - YouTube	아이스,작대기,졸피뎀,얼음,ghb,물뽕,떨,도리도리,허브,필로폰수면제,히로뽕수면제,필로폰	12
16520	작대기,아이스,크리스탈,필로폰작대기,아이스 샘플,떨,작대기판매,필로폰 ...	떨,크리스탈,필로폰작대기,아이스	4
16763	운동의 마약성0722 22시02분 CH9 1 KBS1 - YouTube	뽕,술,졸피뎀	3
17223	가짜 약 제품 아이스 카톡 angel700 작대기 필로폰 판매 - YouTube	술,허브,필로폰크리스탈,작대기,아이스,최상급,물뽕	7
17343	작대기 아이스 팝니다 텔레그램 ID:aassdd123 - YouTube	아이스,졸피뎀,ghb작대기,히로뽕,순도,물뽕,도리,술,최상급,여성흥분제,필로폰,얼음작대기,크리스탈	13
17385	카톡 angel700 ...	떨,크리스탈,ghb	3
17546	All comments on 카톡jayk1 물뽕 수면제GHB 아이스 작대기 ...	작대기,물뽕,ghb아이스,수면제아이스	4

Fig. 6. Collected URL list

시스템을 통해 확인된 웹사이트 중 이전에 작성한 불법·유해정보 게시글은 2012년이며, 아직까지도 삭제되지 않고 존재하여 웹사이트의 관리가 소홀 하다는 것을 확인 할 수 있다. 또한, 이 시스템을 통해 링크된 도메인에 직접 접속하여 보면 검색엔진에 검출되지 않는 불법·유해정보 게시글이 다수 존재하는 것을 확인할 수 있다.

향후 본 논문에서는 사용한 탐지시스템에 구글의 완전한 검색 결과와 다른 검색엔진을 통한 검색 결과를 추가하면 불법·유해정보 탐지 업무를 좀 더 효율성 있게 할 수 있을 것으로 기대 한다.

Fig. 7. Drug sale posts in Youtube

V. 결론 및 향후 연구 방향

본 연구에서 제안하는 불법·유해정보 탐지 시스템은 검색어 사전 기반으로 구글 검색을 통해 웹(Web)상의 공개 정보를 자동으로 수집·분석 할 수 있게 하였다. 자동으로 수집된 웹페이지들은 도메인별로 분류하여 사용자가 몇 번의 동작만으로 불법·유해정보의 페이지를 확인할 수 있는 장점이 있다. 본 연구에서 제안한 시스템은 웹페이지 수집 개수를 기준으로 도메인의 순위를 부여하기 때문에 신규 유해 웹사이트 탐지에 있어 효율적이다. 또한, 이 시스템을 통하여 유해 사이트들을 정보를 시각적으로 빠르게 파악할 수 있어 모니터 요원들의 심의 처리를 진행하는데 많은 도움이 될 것으로 예측한다. 기존 불법·유해정보 탐지를 인력에 의존하였지만 검색어 사전을 이용한 탐지 자동화 시스템을 통해 유해 사이트들을 한눈에 볼 수 있어 인적 자원 소모의 감소를 가져올 수 있을 것으로 기대된다.

향후에는 수집된 URL이 마약 판매 글인지의 자동화 판단할 연구를 수행할 예정이다. 또한, 빈도수가 높은 마약 판매글이 게시되어 있는 사이트를 중심으로 한 지속적인 감시 시스템을 구축하려 한다. 구축한 감시 시스템을 통해 게시글 속에서 수사에 필요한 정보(이메일, 전화번호, IP 등)를 뽑을 수 있는 방안에 대해 연구가 필요하다.

References

[1] Anonymous Declares War On ISIS Following Paris Massacre, The Huffington Post UK, Nov.16.2015 http://goo.gl/BN4uKs

[2] Tekir, S. Open Source Intelligence Analysis: A Methodological Approach. Saarbrucken, Germany: VDM, 2009.

[3] Yoon, Hae Sung, Yun, Min Woo, "Cyberterrorism : Trends and Reponses" Korean Institute of Criminology Research Series (2012): 4-334, 2012

[4] Gerstein, Josh, "NSA: PRISM Stopped NYSE Attack", Politico. Retrieved June 18, 2013.

[5] Meet Prism's little brother: Socmint, wired, Jun.26.2013 http://www.wired.co.uk/news/archive/2013-06/26/socmint

[6] Four hours a day every day, people who watch the porn, JoongAng Daily, Nov.08.2015 http://www.koreadaily.com/news/read.asp?art_id=3803400

[7] Clive Best, "Web Mining for Open Source Intelligence," Information Visualisation, 2008. IV'08. 12th , 2015

[8] Best, C. "Challenges in Open Source Intelligence," Intelligence and Security Informatics Conference (EISIC), 2011 European pp. 58-62, 2011

[9] Neri, F., Pettoni, M., "Stalker, A Multilingual Text Mining Search Engine for Open Source Intelligence," Information Visualisation, 2008. IV '08. 12th International Conference pp. 314-320, 2008

[10] Christopher J.Rhodes, "The Use of Open Source Intelligence in the Construction of Covert Social Networks," Counterterrorism and Open Source Intelligence, 2011

[11] Watters, Paul A., and Nigel Phair. "Detecting illicit drugs on social media using automated social media intelligence analysis (ASMIA)." Cyberspace Safety and Security. Springer Berlin Heidelberg, pp. 66-76, 2012

[12] jsoup: http://jsoup.org/

[13] Kkokkoma lexical analyzer : http://kkma.snu.ac.kr/

[14] Lee, Sungjick, and Han-joon Kim. "News keyword extraction for topic tracking," Networked Computing and Advanced Information Management, 2008. NCM'08. Fourth International Conference on. Vol. 2. IEEE, 2008.

[15] SeleniumIDE: http://seleniumhq.org/projects/ide/

〈저자소개〉

한 병 우 (Byeong-woo Han) 학생회원
20022년 2월: 한림대학교 컴퓨터공학과 졸업
2014년 9월~현재: 고려대학교 정보보호학과 석사과정
〈관심분야〉 정보보호, Cyber Intelligent, OSINT

윤 지 원 (Ji Won Yoon) 종신회원
2003년 2월: 성균관 대학교 정보공학 졸업
2005년 2월: University of Edinburgh, 정보학과 석사 졸업
2008년 11월: University of Cambridge 전자공학과 박사 졸업
2008년 2월~2009년 5월: University of Oxford, 로봇연구소 박사후과정
2011년 7월~2012년 8월: IBM 연구소 정규 연구원
2012년 9월~현재: 고려대학교 정보보호대학원 부교수
〈관심분야〉 신호정보처리, 응용통계, 도감청 탐지기술