

디지털포렌식 공모전 제출 논문

사이버 수사관을 위한 공학도의 크롤러 연구일지

Engineering Student's Crawler Research Journal
for Cybercriminal Investigators

2019년 10월

영산대학교
사이버경찰학과
김 태 룡

목차

국문 요약	vi
영문 요약	vii
제 1 장 서론	1
1.1 연구 배경 및 목적	1
1.2 논문 내용 및 구성	2
제 2 장 관련 연구	3
2.1 유해사이트 특성	3
2.1.1 미디어 활용	3
2.1.2 운영	4
2.1.3 유해사이트 간 연결	5
2.2 유해사이트 유형 및 분석	6
2.2.1 사이버 음란물	6
2.2.2 사이버 도박	8
2.2.3 마약 및 향정신성 의약품	10
2.2.4 총포 및 화약류	11
2.2.5 불법 거래	12
2.2.6 불법 게임물	13
2.3 유해사이트 모니터링 효율	15
2.3.1 직설적 키워드 검색	16
2.3.2 은어 검색	16
2.3.3 구글 검색 옵션 사용	17
2.3.4 유해사이트 내에서 탐색	17
제 3 장 유해사이트 탐색 크롤러	19
3.1 유해사이트 탐색 크롤러 알고리즘	19
3.2 버전별 실험 및 평가	21
3.2.1 실험 환경	21
3.2.2 DragonEye	23
3.2.3 NuriCHAIN beta 1.0.0	24
3.2.4 NuriCHAIN beta 1.0.1	24
3.2.5 NuriCHAIN beta 1.0.2	25
3.2.6 NuriCHAIN beta 1.0.3	26
3.2.7 NuriCHAIN beta 1.0.4	27
3.2.8 NuriCHAIN beta 1 조건부 실험	28

3.2.9 NuriCHAIN beta 2.0.0	30
3.2.10 NuriCHAIN alpha 1.0.0	34
제 4 장 데이터 분석	45
4.1 데이터 수집 알고리즘	45
4.2 데이터 분석	46
4.2.1 CSV 파일 출력	46
4.2.2 Pie Chart 출력	48
4.3 시각화	49
4.3.1 IP주소/URL주소 변경 타임라인	49
4.3.2 유해 사이트 추적 단서 마인드맵	49
제 5 장 결론	51
5.1 효과 및 기대효과	51
5.2 한계 및 향후 연구 과제	52
참고문헌	54

표 목 차

[표 1-1] 논문 흐름도	2
[표 2-1] 직설적 키워드 검색 결과	16
[표 2-2] 은어 검색 결과	16
[표 2-3] 구글 검색 옵션 사용	17
[표 3-1] 유해사이트 탐색 크롤러 동작 순서	19
[표 3-2] DragonEye, NuriCHAIN beta 1버전 실행 환경	22
[표 3-3] NuriCHAIN beta 2버전 실행 환경	22
[표 3-4] NuriCHAIN alpha 1버전 실행 환경	22
[표 3-5] DragonEye 실행 결과	23
[표 3-6] NuriCHAIN beta 1.0.0 실행 결과	24
[표 3-7] NuriCHAIN beta 1.0.1 실행 결과	24
[표 3-8] NuriCHAIN beta 1.0.3 실행 결과	26
[표 3-9] 조건부 실험 환경	29
[표 3-10] 실험 내용	29
[표 3-11] NuriCHAIN beta 2의 업데이트 내용	31
[표 3-12] NuriCHAIN beta 2의 모니터링 결과	33
[표 5-1] NuriCHAIN 앱 구동 알고리즘	53

그림 목 차

[그림 2-1] SNS 및 게시판을 이용한 유해사이트 홍보	3
[그림 2-2] URL과 디자인만 바꾼 동일한 범죄조직의 두 사이트	4
[그림 2-3] 유해사이트 운영의 기본 구조	5
[그림 2-4] 각기 다른 도박 사이트 광고 배너	5
[그림 2-5] 실제 운영 중인 유해 사이트의 소스코드 속 SNS주소	6
[그림 2-6] 유해 매체 제공 사이트와 SNS계정 간 공유 관계도	7
[그림 2-7] 실제 복제 계정들 간 리블로그 내역	7
[그림 2-8] 홍보 게시글에서 추출한 사진	8
[그림 2-9] 불법 도박 사이트 관계도	9
[그림 2-10] 넘쳐나는 계정과 복제 사이트	9
[그림 2-11] 활동 닉네임 단위로 저장된 데이터	10
[그림 2-12] 유해 키워드가 저장된 엑셀 파일	11
[그림 2-13] 총기 규제가 없는 국가의 언어의 경우, 검색 결과가 많은 모습	12
[그림 2-14] 일반 과세자로 등록된 휴대폰 갯 업체	13
[그림 2-15] 국가법령정보센터 정보통신망법 검색 결과	13
[그림 2-16] 3개월간 운영하고 새로 열겠다는 의지를 대놓고 담은 공지	14
[그림 2-17] 복제 사이트에 디스코드를 통한 회원 관리에 전자상거래까지	15
[그림 2-18] 유해 매체는 링크를 타고	18
[그림 3-1] 유해사이트 탐색 알고리즘	20
[그림 3-2] 필자가 만들 크롤러는 렌더링을 하지 않는 방식을 채택하였습니다 ·	21
[그림 3-3] DragonEye 실행 화면	23
[그림 3-4] DragonEye 출력 로그	23
[그림 3-5] 키워드 버튼이 생긴 모습	25
[그림 3-6] CSV 형식 로그 출력 모습	25
[그림 3-7] 추천 키워드 요청버튼 추가	26
[그림 3-8] GUI 변경	27
[그림 3-9] 버전 확인 페이지 개설	27
[그림 3-10] 키워드 및 URL 요청 화면	27
[그림 3-11] 키워드 타입 설정 창	27
[그림 3-12] NuriCHAIN beta 1 성능 분석 차트	28
[그림 3-13] 조건부 실험 결과 차트	29
[그림 3-14] NuriCHAIN beta 2의 메인 화면	32
[그림 3-15] 웹 페이지에서 제공하는 통계 그래프	33

[그림 3-16] maximumDepth, maximumEmptyHand로 성능을 향상시켰다	34
[그림 3-17] 베스트 누리캡스 상장 및 트로피	35
[그림 3-18] 누리캡스 회원 인증 방식	36
[그림 3-19] 동료 개발자에게 온 의뢰서 중 일부	37
[그림 3-20] 클라이언트와 서버 개발 환경	38
[그림 3-21] 게임 메뉴 바 형식 메인 메뉴 GUI	39
[그림 3-22] 탐색 GUI	39
[그림 3-23] 통계 GUI	40
[그림 3-24] 자동 서식 생성 GUI	41
[그림 3-25] 도움말 GUI	41
[그림 3-26] 임의의 유해 게시글을 발굴한 모습	42
[그림 3-27] 검색 결과를 CSV로 출력한 모습	43
[그림 3-28] NuriCHAIN 4월 사용통계	43
[그림 4-1] 데이터 추출 알고리즘	45
[그림 4-2] NuriCHAIN 크롤링 생태계	46
[그림 4-3] 수집된 유해사이트검색 결과	47
[그림 4-4] 다운로드 화면	47
[그림 4-5] 다운로드 된 CSV 파일	47
[그림 4-6] 다운로드 된 테스트용 키워드 파이 차트	48
[그림 4-7] 타임라인 형식으로 펼쳐진 주소 변동 내역	49
[그림 4-8] 아이콘과 텍스트로 이루어진 트리 시각화	50
[그림 5-1] NuriCHAIN 피드백	51

국문 요약

사이버 수사관을 위한 공학도의 크롤러 연구일지

김태룡
사이버경찰학과
영산대학교

증가하는 불법 유해정보에 적극 대응하기 위해 사이버안전국은 2007년부터 경찰과 시민으로 결성된 사이버 명예경찰 집단인 누리캡스(NuriCops)를 창설하였습니다. 필자는 누리캡스 일원으로써 2014년부터 현재 까지 활동하고 있으며, 주로 유해 사이트 모니터링 및 신고활동을 하고 있습니다.

모니터링활동 중, 유해사이트는 패턴과 유행을 가지는 양상을 확인하여 ‘유해사이트는 패턴을 가졌다’를 전제로 2016년 여름부터 유해사이트 탐색 크롤러(Crawler)를 제작하게 되었습니다.

크롤러를 통해 수집 된 유해사이트를 직접 방문하여 유해 여부를 확인한 결과, 많은 키워드가 걸린 사이트일수록 높은 확률로 유해 사이트임을 확인하였으며, 이후 반복적인 실험을 진행하여 크롤러 업그레이드 및 키워드를 수집 해 나갔습니다.

2019년 봄, 일반 누리캡스 회원들에게 크롤러를 배포한 이후, 전국 누리캡스들의 크롤링 데이터를 수집, 빅데이터를 활용하여 시각화 및 CSV 파일 형태로 출력하는 기능을 추가하여 마약 매매, 음란물 유포, 자살 유도 등의 유해 집단의 추적 단서들을 한 눈에 볼 수 있도록 하였습니다.

누리캡스 및 수사관들을 위한 크롤러 ‘NuriCHAIN’의 3년간 기록이 수사관 및 관계자 혹은 자라나는 프로그래머 및 독자 분들에게 도움이 되었으면 합니다.

ABSTRACT

Engineering Student's Crawler Research Journal for Cybercriminal Investigators

KIM, TAE-RYONG
Cyber Police Department
Yongsan University

To respond to the increasing illegal harmful information, the Cyber Bureau has created NuriCops group with police and citizens in 2007. I have been with NuriCops since 2014. And mainly monitoring and reporting harmful sites.

During the monitoring activities, I noticed that there are patterns and trends on harmful sites. And since the summer of 2016, I developed harmful site search crawlers on the premise that "harmful sites have patterns."

After visiting harmful sites collected through crawler, I found that sites with many keywords are more likely to be harmful. Then repeated the experiment to upgrade and collect keywords.

In the spring of 2019, I distributed the crawler to all Nuricops members and collected various big data. Then I visualized the big data, and output it as a CSV file for the harmful clues found (drug trafficking, porn distribution and suicide).

I hope that the three-year record of NuriCHAIN, a crawler for Nuricops and investigators, will be helpful to investigators and growing programmers and readers.

키워드

C

CPTED

ㄱ

가시화

L

누리캡스

ㅂ

빅데이터

ㅅ

수사

시각화

ㅇ

유해사이트

ㅋ

크롤러

ㅌ

타임라인

ㅍ

파이차트

프로그램

제 1 장 서론

1.1 연구 배경 및 목적

누리caps 회원으로써 유해사이트 모니터링 및 신고 활동을 하던 도중, 유해사이트는 유사한 패턴과 디자인 및 자료들을 가지고 운영되는 것을 알게 되었습니다. 이는 사이트 방문자가 성인이 아닐 경우에도 성인 자료를 보고, 동일한 악성 스크립트 위험에 노출됨을 뜻합니다.

과학기술정보통신부와 한국인터넷진흥원이 공개한 '2018년 인터넷 이용실태조사 결과'에 따르면 대한민국 1975만 가구 중, 99.5%가 인터넷 접속이 가능하고, 청소년의 인터넷 이용률이 99.9%에 달하며, 94.9%의 국민이 스마트기기를 보유하고 있습니다.[1] 이는 누구든 언제 어디서나 유해사이트와 마주할 수 있음을 의미하기에 위의 염려가 현실이 될 가능성이 높습니다.

때문에 당시 더 많은 유해 미디어를 잡기 위해서 유해사이트 모니터링 설명회를 가지고 발표를 하던 도중, '제 2회 SUA 영남지부 컨퍼런스 (2016/04/03)'에 참여하게 되어 창원대학교 Casper 홍성진 발표자의 'Maltego를 이용한 OSINT' 강연을 듣고, '크롤러'를 알게 되었습니다.

크롤러를 이용하여 유해사이트 수집을 자동화 시킬 경우, 사이버 수사대와 누리caps 회원들이 손수 사이트를 방문하지 않더라도 해당 사이트의 유해성을 판단하여 악성 스크립트 감염 위험도 덜고, 탐색 효율도 극대화 시킬 수 있을 것이라 생각하고, 누리caps 회원 및 수사관을 위한 크롤러, NuriCHAIN 제작에 돌입하게 되었습니다.

5년간 6만 여건의 모니터링 및 신고활동을 통해 축적된 노하우로 만들어진 NuriCHAIN이 수사관 및 누리caps 회원들의 활동을 지원(activity support)하여 더 많은 유해 미디어를 제거하도록 돕고, SNS 상의 유해 게시글을 효과적으로 발굴하여 SNS 공간을 깨끗하게 함으로써 (territoriality), 사이버 환경을 변화 시켜 전반적인 사이버 범죄 가능성을 줄이는 것이 목적입니다.[2]

1.2 논문 내용 및 구성

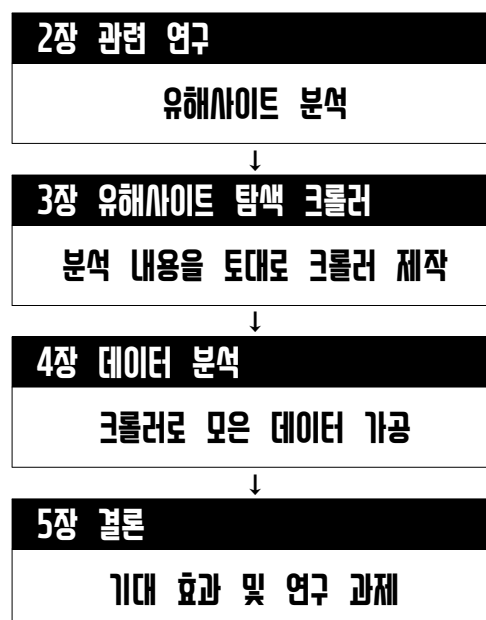
본 논문은 유해사이트 탐색 크롤러 NuriCHAIN의 실험 과정 및 결과를 중심으로 작성되었습니다.

2장에서는 직접 모니터링 및 신고활동을 하면서 체득한 각종 유해사이트에 대한 분석 글을 시작으로, 유해 사이트 간 연결 관계 및 유해사이트의 특성을 이용한 Google 검색 실험을 진행 할 것입니다.

3장에서는 2장에서 정리된 내용을 토대로 제작된 유해사이트 탐색 크롤러 NuriCHAIN에 대한 간단한 설명과 버전별 성능 및 실험 결과에 대해 살펴 볼 것입니다.

4장에서는 NuriCHAIN을 통해 모은 데이터를 각종 결과물들로 추출해 보고, 활용 방안에 대하여 알아볼 것입니다.

5장에서는 기대 효과 및 한계와 향후 연구 과제에 대하여 알아보고 끝을 맺습니다.



[표 1-1] 논문 흐름도

제 2 장 관련 연구

2.1 유해사이트 특성

2.1.1 미디어 활용

유해사이트는 대개 청소년에게 유해한 음란물, 마약 매매, 불법 도박 등의 자료로 구성되어 있으며, 배너 광고 및 물품/자료 판매 등을 통해 수익을 얻습니다. 하지만 검색엔진의 필터링 때문에 검색을 통한 일반인의 접근을 끌어내기란 쉽지 않습니다. 때문에 고객을 유치 위해 대부분 SNS를 활용하거나 관리가 허술한 게시판들 통해 유해사이트 주소가 적힌 광고지를 뿌려 호객 행위를 합니다.



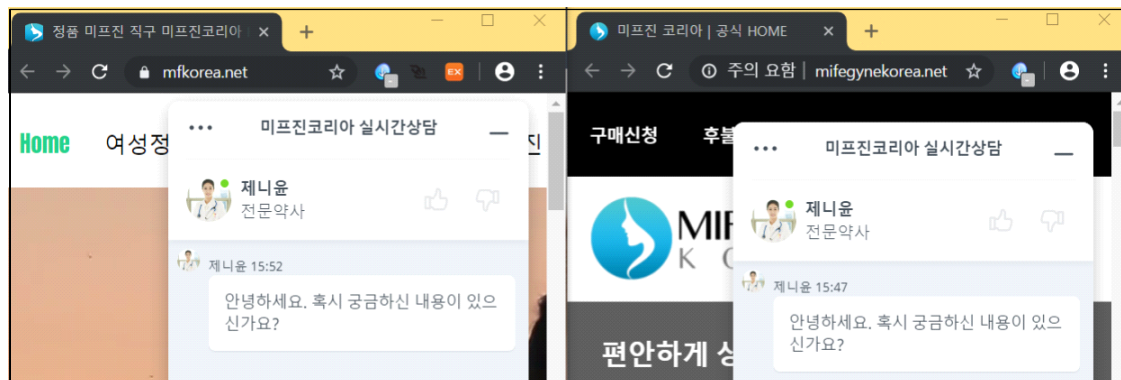
[그림 2-1] SNS 및 게시판을 이용한 유해사이트 홍보

[그림2-1]과 같이 호객행위 시, 접근성은 높이고 필터링은 피하기 위해 판매하려는 물건 이름을 본문에 쓰지 않고 태그로 설정하거나, 특수기호 처리하여 필터링을 피해가는 수법을 사용하여 유해 미디어가 SNS에 버젓이 게시되어 있는 모습을 심심치 않게 볼 수 있습니다.

특히 유명한 SNS 혹은 사이트일수록 유해 미디어가 자주 노출되는데, 인지도가 높은 사이트에서 광고하는 것이 더 많은 불특정 다수가 열람하는 것은 물론, 차단된다 하더라도 사이트 자체가 폐쇄될 일은 없고, 작성된 홍보물 중, 혹은 수많은 계정 중 ‘단 하나’만 제거되는 수준에서 그치기 때문입니다. 그 중에서도 SNS는 계정이 제거되더라도 ‘공유’ 버튼을 통해 공유된 모든 글까지 제거하지 않는다면 언제든 증식이 가능하기 때문에 범죄자들의 주 활동지가 됩니다.

2.1.2 운영

신고 활동을 진행하다보면, A사이트의 내용물이 B사이트의 내용물과 동일한 경우가 빈번합니다. 이는 하나의 범죄조직이 여러 복제 사이트를 개설 한 것으로, 웹 호스팅을 이용한 단기 운영 사이트입니다.

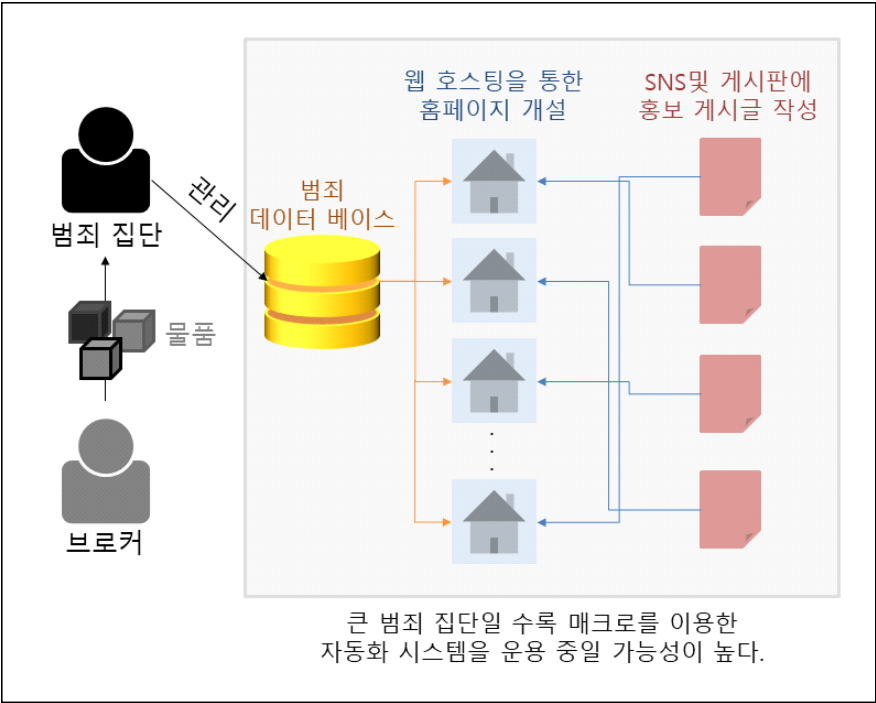


[그림 2-2] URL과 디자인만 바꾼 동일한 범죄조직의 두 사이트

무료 웹 호스팅을 이용하기에 대부분 30일 이내에 사라지며, 신고를 하여 차단시키더라도 새로운 주소로 계속해서 생성되는 특징을 가집니다. 이처럼 SNS 또한 계정을 계속해서 생성할 수 있기 때문에 신고를 하여도 직접적인 타격이 없지만, 신고를 하지 않는다면 그 사이에 피해자가 생길 수 있으므로 일개 게시글 하나를 발견하더라도 놓치지 않고 신고 해 주는 것이 중요했습니다.

추 후 각 유해사이트를 유형별로 자세히 알아볼 것이지만, 대부분의

유해사이트 운영은 아래 [그림 2-3]의 구조를 바탕으로 이루어집니다.



[그림 2-3] 유해사이트 운영의 기본 구조

2.1.3 유해사이트 간 연결

유해사이트를 한 번이라도 접속 해 보신 독자 분들은 아마 몇 번은 보셨으리라 생각합니다. “분명 나는 A 도박 사이트에 접속 했는데, 왜 B 도박사이트와 C 도박 사이트까지 선전하고 있지?”



[그림 2-4] 각기 다른 도박 사이트 광고 배너

이는 각기 다른 범죄 조직이 운영 중인 사이트보다도, 동일한 하나의 조직이 운영 중인 ‘이름만 바꾼’ 복제 사이트일 가능성이 높습니다. 교토삼굴 [狡兎三窟] (영리한 토끼는 세 개의 굴을 판다) 이라는 말이 있듯이, A 도박 사이트가 폐쇄 되어도 이용자가 광고를 타고 들어간 B, C 사이트를 즐겨찾기 했다면 수입에 지장이 없기 때문입니다.

후술할 도박 사이트 운영 시스템 때문에 복제 사이트 홍보는 도박 사이트에서 가장 도드라지게 나타나는 특징이지만, 가끔 도박 이외의 유해 매체가 복제 사이트 홍보를 통해 대피소(차단당할 시, 임시로 운영되는 사이트) 운영 주소를 사용자들에게 흘리는 모습도 관찰됩니다.

2.2 유해사이트 유형 및 분석

2.2.1 사이버 음란물

음란물의 경우 웹 사이트 및 웹 하드보다 SNS(Tumblr, Twitter)를 통한 접근성이 높기 때문에 SNS 상으로 많이 노출되는 편입니다. 또한 음란물 동영상 스트리밍 사이트의 경우, 서버가 발각되어도 자료가 남지 않도록 SNS 상에 올라온 동영상을 끌어다 오는 형태로 점점 변화하고 있기 때문에, SNS상에 올라오는 음란물을 신고하는 방법도 사이트 하나를 잡는 만큼의 가치가 있었습니다.

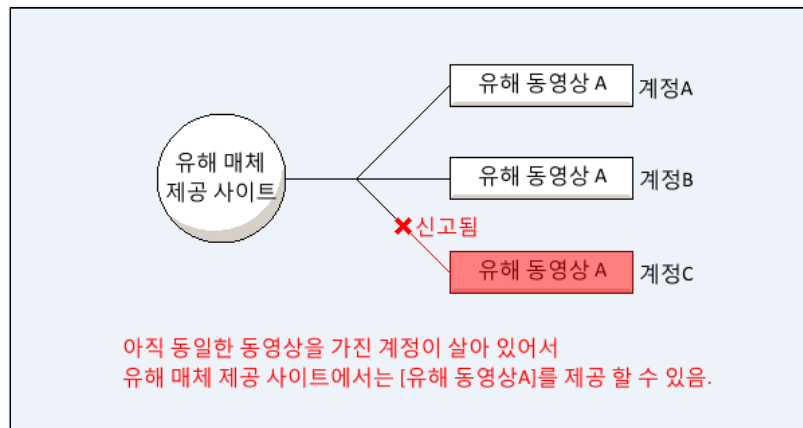
```

136 <section id="bo_v_atc">
137 <h2 id="bo_v_atc_title">본문</h2>
138
139
140 <div class="phone"> <p><a href="tel:060-1111-1111" id="phone">대화 와 셉/892 가 일
141 </div>
142 <div class="tag">
143 <p>서양, 로또, 최강방문화, 비아, 서울레이스, 긴발플러스, 낭랑18세, 스왐, VIP멤버십 ,여배우야 뽕</p>
144 </div>
145 <video src="https://v.tumblr.com/tumblr_o6_1je7ug.mp4" width="100%" he
146
147
148 <div class="view_bn">
149
150 <a href="http://.com/" target="_blank"><img src="http://.com/spV4ybr
152 </div>
153 </section>
154 </article>
155
156 <script>

```

[그림 2-5] 실제 운영 중인 유해 사이트의 소스코드 속 SNS주소

SNS에 유포되는 불법 유해 매체의 경우, 계정 하나가 차단당할 경우, 포스팅 했던 모든 게시글이 차단당하기 때문에 복제 계정을 여럿 두고, 게시글을 리블로그(자신의 계정으로 복사)하는 형태가 주를 이루고 있습니다. 또한 이렇게 복제된 계정들을 유해 매체 제공 사이트에서 스트리밍하게 될 경우, 아래[그림 2-6]과 같이 SNS계정 하나를 잡더라도, 동일한 게시글을 가진 계정이 하나라도 남아 있다면 타격을 입히지 못했습니다.



[그림 2-6] 유해 매체 제공 사이트와 SNS계정 간 공유 관계도

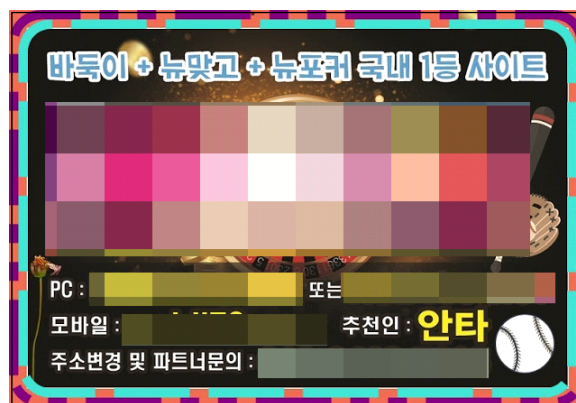
따라서 음란물 관련 유해 매체 신고 시에는 리블로그한 계정까지 모두 신고하는 것이 좋으며, [페이지 소스보기] 기능을 이용하여 동영상/사진의 절대 경로 URL을 신고하는 것이 가장 좋았습니다.



[그림 2-7] 실제 복제 계정들 간 리블로그 내역

2.2.2 사이버 도박

인터넷에 퍼져있는 사이버 도박의 경우, [도박 사이트]와 [홍보 게시글]로 분류할 수 있습니다. 이들은 단속을 피하기 위해 ‘가입 코드’(추천인)를 입력 해야만 회원가입이 되도록 운영 중이며, 홍보 게시글에서 어렵지 않게 가입 코드와 도박장 URL을 얻을 수 있습니다.

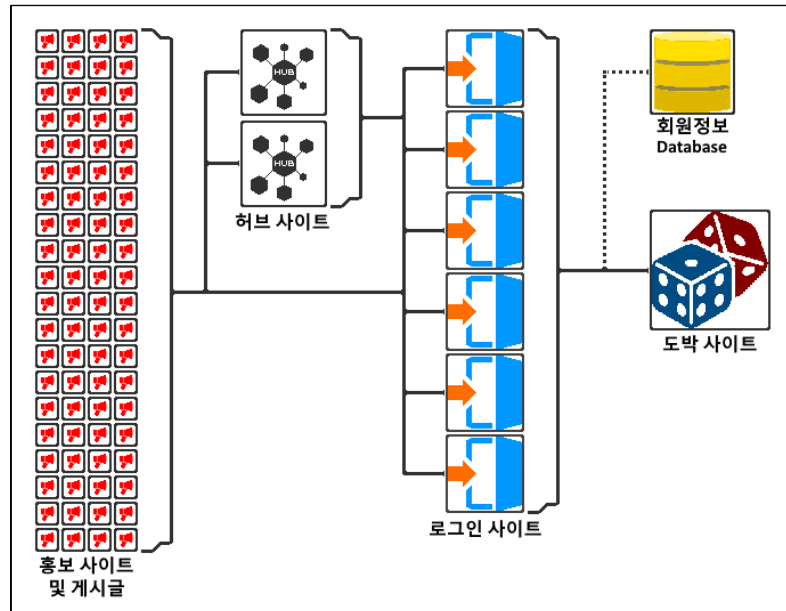


[그림 2-8] 홍보 게시글에서 추출한 사진

로그인 사이트를 다른 누리caps 회원이 먼저 신고 한 경우는 수 없이 봐 왔지만, 회원가입 이후 넘어가는 게임 사이트가 신고 된 경우는 매우 적었습니다. 또한, 각기 다른 로그인 사이트에서 동일한 회원정보 DB를 사용하여, A 사이트에서 회원가입 해도 B 사이트에서 로그인이 가능하였으며, 로그인 이후 동일한 게임 사이트로 연결되었습니다.

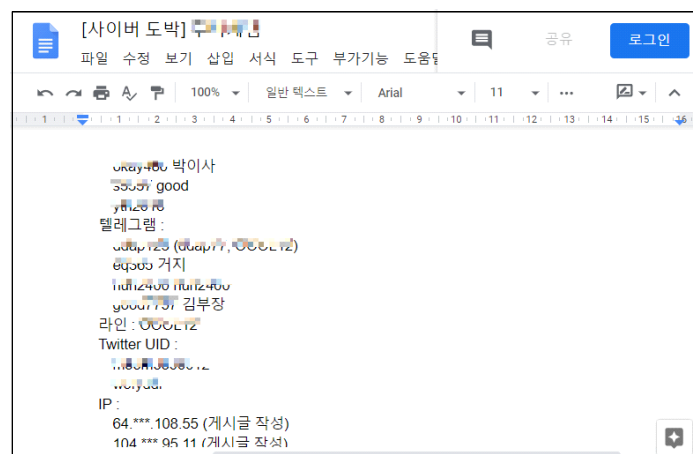
즉 도박 사이트 관리자는 로그인 사이트와 게임 사이트를 분리하여 깊이 모니터링 하지 않으면 게임 사이트의 존재를 모르도록 만듦으로써, 신고 시 로그인 사이트만 폐쇄 당하되 게임 사이트를 살리려는 의도를 알 수 있었습니다.

그 외에도 각종 도박 사이트들의 URL과 홍보글로만 가득한 허브사이트를 중간 중간 개설하여, 차단된 로그인 사이트의 접속 경로를 수시로 업데이트 해 주는 치밀함 까지 보여 주었습니다.



[그림 2-9] 불법 도박 사이트 관계도

이러한 특징 때문에 추적 단서들을 웹 문서에 저장하다 보니, 한 눈에 보기 좋아지는 것을 알게 되었습니다. 이는 후에 서술할 데이터 가시화 프로그램, NuReport 제작의 초석이 되었습니다.

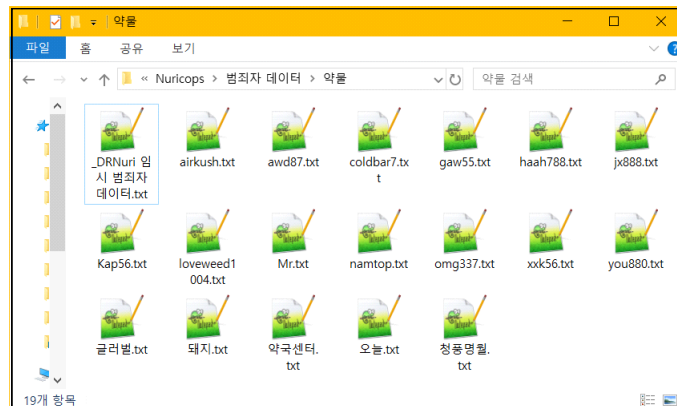


[그림 2-10] 넘쳐나는 계정과 복제 사이트 1)

1) https://docs.google.com/document/d/1PmXrm_zQuk-2VheMytXJy27jba3tjh2rcOhiDqgPUvk

2.2.3 마약 및 향정신성 의약품

마약 및 향정신성 의약품 판매의 경우, 실질적인 물품이 오가는 매매이기 때문에 반드시 거래자의 연락처가 노출되었습니다. 하지만 잦은 모니터링 활동으로 인해 매매상이 연락처를 수시로 변경하여 추적을 피하므로, 특별한 경우를 제외하고는 매매상이 사용하는 ‘닉네임’ 혹은 ‘상호’를 통해서 동일범 여부를 판단하였습니다.



[그림 2-11] 활동 닉네임 단위로 저장된 데이터

실제 상품이 오가는 마약 매매상은 닉네임에 신뢰도를 걸고 활동하는 집단이기 때문에, 오랜 기간 동안 닉네임을 유지 해 온 매매상일수록, 연락처가 자주 바뀌지 않은 매매상일수록, 취급 품목이 여러 가지가 아닌 매매상일수록 사기꾼이 아닐 가능성이 높았으며,(댓글에 사기 당했다는 내용이 그렇지 않은 집단에 비해 적게 발견되었습니다.) 향후에도 활동할 가능성이 높았습니다. 때문에 마약 매매상의 게시글은 다른 유해 게시글을 찾는 것 보다 쉬웠습니다. 그들이 사용하는 닉네임만 검색하면 그간 활동한 게시글이 모두 나왔기 때문입니다.

다만, 마약류는 다른 범죄 행위들에 비해 은어가 굉장히 많기 때문에 은어와 그 지칭 대상을 모른다면 잡지 못할 뿐만 아니라, 은어와 상품이 해마다 새로운 이름과 상품으로 변경되기 때문에, 주기적인 모니터링을 하지 않을 경우, 노련한 누리caps 회원이라도 놓칠 가능성이 높

아 보였습니다.

때문에 마약 및 향정신성 의약품은 물론, 음란물 및 다양한 유해매체에서 사용되는 은어 및 키워드를 정리한 문서를 만들어 7월에 누리캡스 홈페이지에 게시하였습니다. 이후 8, 9월 누리캡스 전체 신고 건수를 보니, 6만 6천 건으로, 15년도 전체 누리캡스 신고량(3만7146건)의 약 2배가량 높게 나온 것을 확인하였습니다.[3]

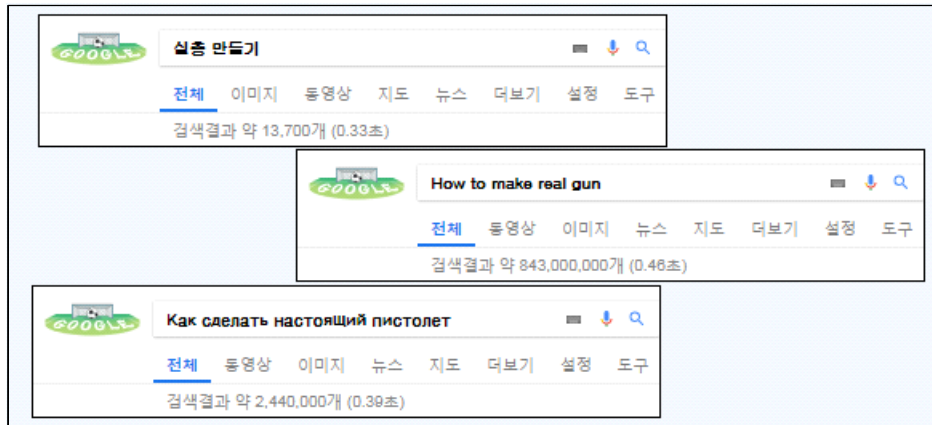
	A	B	C	D	E	F
63	마약류	잡초	마약	제한없음	약물	가짜 대마초
64	마약류	최음제	마약	일부 필터링	약물	메스암페타민/GH8 등. 최음 효과를 내는 마약을 주로 사용함.
65	마약류	케타민	마약	제한없음	약물	의료용 마취제(진통제). 오/남용시 정신성 마약.
66	마약류	코카인	마약	일부 필터링	약물	신경계 작용 마약. 의학용도 외 소지는 불법.
67	마약류	쿠쉬	마약	제한없음	약물	대마초 품종 중 하나.
68	마약류	크랙	마약	제한없음	약물	코카인+베이킹파우더 (흡연용)
69	마약류	크리스탈	마약	제한없음	약물	결정 형태의 질 좋은 메스암페타민
70	마약류	필로폰	마약	일부 필터링	약물	필로폰, 히로뽕이란 명칭은 일본에서 합법적으로 판매하던 상품
71	마약류	해시시	마약	제한없음	약물	대마초를 고체로 굳힌것.
72	마약류	헤로인	마약	일부 필터링	약물	아편 성분 중 하나인 모르핀을 정제하여 얻는 것으로, 모르핀보
73	마약류	해인즈	마약	제한없음	약물	대마초 품종 중 하나.
74	마약류	히로뽕	마약	일부 필터링	약물	필로폰의 은어. 즉 메스암페타민.
75	마약류	고농축	마약 용어	제한없음	약물	장사 멘트중 하나. 마약의 품질이 좋다는 것을 어필함.
76	마약류	던지기	마약 용어	제한없음	사기	지정된 장소에 물건을 두고 간다는 거래 방식.
77	마약류	드랍	마약 용어	제한없음	사기	지정된 장소에 물건을 두고 간다는 거래 방식.

[그림 2-12] 유해 키워드가 저장된 엑셀 파일

이는 엄선된 키워드를 이용하여 모니터링 할 경우, 높은 효율을 기대할 수 있음을 보여주는 자료로, 후에 서술할 유해사이트 탐색 크롤러 NuriCHAIN의 키워드 품질 개선 알고리즘에 영향을 주게 됩니다.

2.2.4 총포 및 화약류

총포 및 화약류 매매 및 제조법의 경우, 국내에서 강한 규제를 하고 있기 때문에 포털 사이트나 각종 국내 게시판에서는 거의 나오지 않습니다. 때문에 범죄 의향이 있는 사람의 경우, 해외 문서로 눈을 돌릴 것이라 판단하여 총기 규제가 없는 국가의 언어로 검색을 시도 해 보았더니, 국내 검색보다 총포 및 화약류 제작법이 많이 나왔습니다.



[그림 2-13] 총기 규제가 없는 국가의 언어의 경우, 검색 결과가 많은 모습

총기 및 폭발물은 화약이 주체이기 때문에, 화약의 재료가 되는 물질(KNO₃)을 추출하는 법에 대해 검색할 경우, 많은 유해 정보를 수집할 수 있었으며, 총기라고 꼭 화약이 필요하지 않고, 폭탄이라고 꼭 파편이 튀는 법이 없듯 코일건, 레일건, 화염방사기, 섬광탄 등 다양한 종류의 총기 및 폭발물 제조법 또한 단속에 걸리지 않고 버젓이 게시되어 있었습니다. 이 덕에 불법 총기류 미디어를 잡기 위해서는 총기의 종류를 넓게 잡고, 다양한 언어를 사용하여 모니터링을 해야 한다는 것을 알게 되었습니다.

2.2.5 불법 거래

불법 명의 거래의 경우, 과거에는 통장 매매, 신분증 복제 등을 빌미로 실질적인 개인정보가 담긴 물품이 오갔지만, 현재는 비트코인 운반 알바, 카드 깡 등 신종 수단을 이용하는 추세가 늘어나고 있습니다.

특히 경기가 나쁠 때는 ‘고수익 알바’와 같이 달콤한 말로 자금 세탁책을 모집하는 경우도 많이 보였으며, 휴대폰 깡(소액 결제 현금화 등)과 같이 다양한 수단으로 고리대금업을 펼치는 모습에, 휴대폰깡 피해 사례까지 생겨[4] 눈에 거슬렸습니다.

모니터링 결과, 불법 거래 또한 타 유해사이트와 마찬가지로 무료 웹 호스팅을 이용한 차단되어도 타격 없는 형태로 운영되고 있었습니다.

한 가지 다른 점은, 사업자 등록번호를 등록하고 운영하는 업체가 대부분에, 웹페이지 하단에 사업자정보(상호, 사업자번호, 사업장 등)를 공개해 놓아, 이것이 진짜 유해사이트인지 일반 사업용 사이트인지 구별이 힘들다는 것이었습니다.

사업자등록상태조회	
사업자등록번호	사업자등록상태
270-43-00000	부가가치세 일반과세자 입니다.

[그림 2-14] 일반 과세자로 등록된 휴대폰 갱 업체

때문에 국가법령정보센터를 뒤져 보았으며, 그 결과 현행법상 정보통신망을 이용하여 재화를 판매하는 행위가 불법임을 찾아내어 모니터링 및 신고활동을 하기 시작하였습니다.

정보통신망 이용촉진 및 정보보호 등에 관한 법률 (약칭: 정보통신망법)	
[시행 2019. 6. 25.] [법률 제16021호, 2018. 12. 24., 일부개정]	
[과] [연] 제72조(벌칙) ① 다음 각 호의 어느 하나에 해당하는 자는 3년 이하의 징역 또는 3천만원 이하의 벌금에 처한다. <개정 2015. 1. 20., 2015. 3. 27.>	
1. 삭제 <2016. 3. 22.>	
2. 제49조의2제1항을 위반하여 다른 사람의 개인정보를 수집한 자	
2의2. 「재난 및 안전관리 기본법」 제14조제1항에 따른 대규모 재난 상황을 이용하여 제50조의8을 위반하여 광고성 정보를 전송한 자	
3. 제53조제1항에 따른 등록을 하지 아니하고 그 업무를 수행한 자	
4. 다음 각 목의 어느 하나에 해당하는 행위를 통하여 자금을 융통하여 준 자 또는 이를 알선·중개·권유·광고한 자	
가. 재화등의 판매·제공을 가장하거나 실제 매출금액을 초과하여 통신과금서비스에 의한 거래를 하거나 이를 대행하게 하는 행위	

[그림 2-15] 국가법령정보센터 정보통신망법 검색 결과

불법 거래의 경우, 일단 위법성 여부만 확인이 될 경우, 모니터링은 어렵지 않았습니다. 재화가 오고가는 거래답게 대포폰, 대포통장일지라도 반드시 정확한 번호가 기재되었기 때문입니다. 이러한 번호들과 사업자 정보들을 모아 트리를 구성하니, 하나의 범죄그룹이 나왔으며, 이는 후에 서술할 데이터 가시화 프로그램, NuReport의 시각화 알고리즘의 밑바탕이 되었습니다.

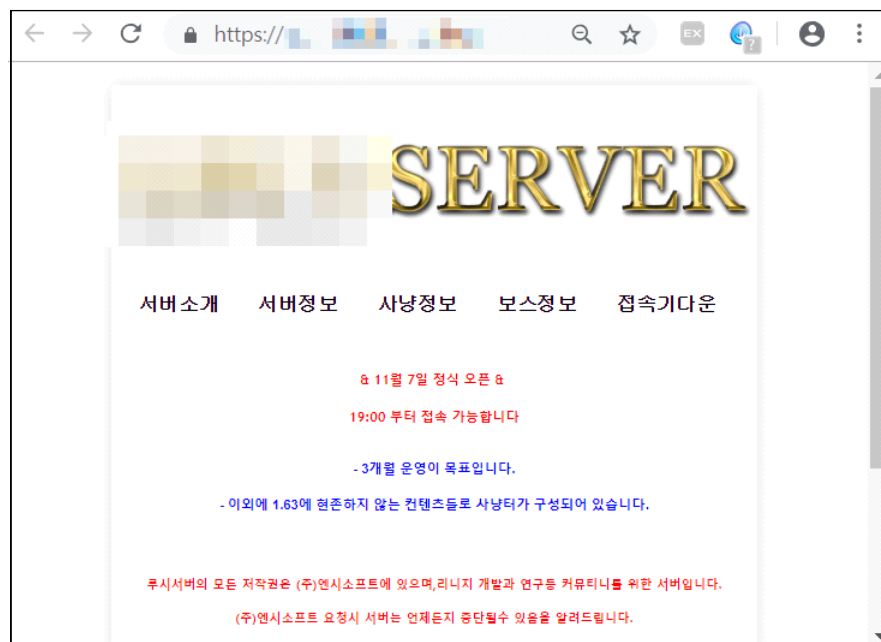
2.2.6 불법 게임물

불법 사설 서버, 게임 핵, 게임머니 환전 등이 불법 게임물에 해당

하며, 이들은 ‘게임’을 대상으로 하기에 불법 게임물 모니터링 시, 아무 주제 없이 찾고 다니는 것 보다는 평소 즐겨 하던 게임 이름을 입력하여 찾는 것이 효율이 높았습니다.

불법 사설 서버의 경우 저작권이 엄연히 게임 회사에 있음에도 불구하고 개인적인 서버를 열어 운영하는 것으로, 프리메이플, 프리바람 등 각종 형태로 운영 중인 사설 서버를 뜻하며, 이들이 그저 즐기기 위한 목적으로 사설 서버를 열었다면 게임 회사의 피해는 있을지언정 유저들의 피해는 없겠지만, 유저들을 범행 대상으로 삼은 사설 서버가 대부분이었습니다.

유저들의 현금 결제를 유도한 후, 일정 액수가 모이면 서버를 종료하고 새 이름, 새 시즌으로 오픈하는 등, 옛 게임의 향수를 악용하는 운영 기법이었는데, 고전 향수가 강한 게임일수록(리니지, 바람의나라 등) 더 많은 불법 사설 서버가 발견되었습니다.



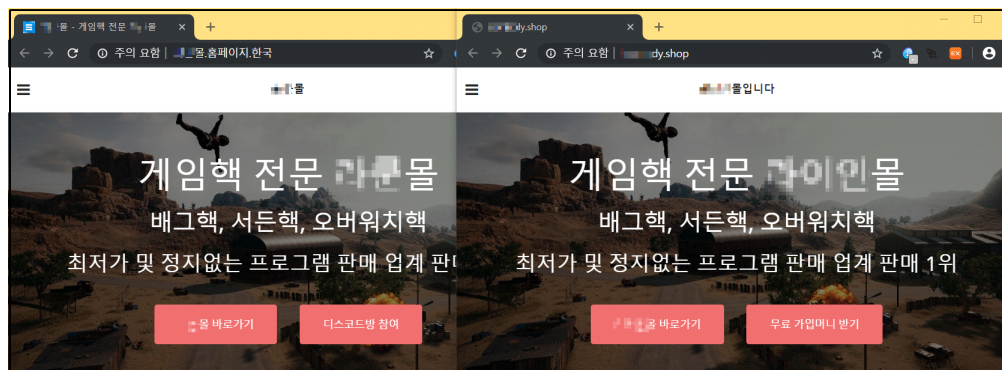
[그림 2-16] 3개월간 운영하고 새로 열겠다는 의지를 대놓고 담은 공지

뿐만 아니라 사설 서버의 경우, 접속기를 따로 설치해야 하는데, 접

속기 설치 시 백신 프로그램을 끄고 설치하라는 등 척 봐도 이상한 권 모술수를 부리며 악성 프로그램을 설치하려 드는 곳이 많았습니다. 이를 가상환경(VMware)에서 도박 사이트와 마찬가지로 직접 해 본 결과, 회원가입, 로그인 까지는 순탄하게 흘러가다가, 게임 설치 시 대부분 바이러스가 동작하였습니다.

이 외에도 불법 사설 서버 모니터링 시, 하자/프리/놀자 등의 키워드를 함께 섞어 줄 경우, 검색 결과가 더욱 풍성해 지는 것을 확인하였으며, 사설 서버를 모아 두는 허브 사이트가 곳곳에 운영되고 있는 모습을 직접 보게 되었습니다.

게임 해킹의 경우, 정당한 게임 플레이를 저해하는 프로그램으로, 개발자가 만드는 것이기에, 다른 불법 유해 매체 운영자들보다 사이버 공간에서 영리하게 추적을 피해 다니는 모습을 보여주었습니다. (Discord 비밀 그룹 채팅방 개설 및 역공학 방지 코드 작성 등) 따라서 ‘개발자를 잡는다’는 생각 보다는 ‘장사를 계속 방해하자’라는 생각으로 꾸준하고 느긋하게 해킹 판매 게시글이 올라 올 때 마다 신고를 하였습니다.



[그림 2-17] 복제 사이트에 디스코드를 통한 회원 관리에 전자상거래까지

2.3 유해사이트 모니터링 효율

2.3.1 직설적 키워드 검색

필자는 모니터링 시, Google 검색을 주로 사용하였습니다. 특히 누리캡스 초창기 시절, 은어도 몰랐던 필자는 대중들에게 잘 알려진 있는 그대로의 단어를 사용하여 모니터링을 하였습니다.

검색	검색 결과	유해성 (1 페이지 기준)
대마 삽니다	약 1,610,000개	매우 높음
엑스터시 삽니다	약 22,600개	매우 높음
게임머니 환전	약 5,670,000개	낮음
명 의 거래소	약 3,090,000개	매우 낮음
섹스 파트너	약 3,120,000개	보통

[표 2-1] 직설적 키워드 검색 결과

있는 그대로의 단어를 사용했더니, Google 측에서도 성인인증을 하라는 안내 문구가 나왔으며, 마약류를 제외한 대부분의 유해 키워드는 검색 결과 1페이지 안에 5건 미만의 유해 게시글이 발견되었습니다.

2.3.2 은어 검색

앞서 검색한 키워드를 동일한 뜻을 가진 은어로 교체한 후, 재검색을 시도하였습니다.

검색	검색 결과	유해성 (1 페이지 기준)
뽕 삽니다	약 2,570,000개	매우 높음
몰리 삽니다	약 722,000개	매우 높음
머니상	약 12,000,000개	매우 높음
장집	약 31,300개	매우 높음
섹파	약 3,510,000개	매우 높음

[표 2-2] 은어 검색 결과

직설적 키워드를 사용했을 때 보다 대부분의 검색 결과가 풍성해졌으며, 유해게시글 또한 1페이지 내 8건 이상 나오는 등, 확연한 차이를 보여주었습니다. 특히 장집, 머니상의 경우, Google의 위험 경고를

문구가 나타나지 않는 모습을 확인할 수 있었습니다.

2.3.3 구글 검색 옵션 사용

앞서 사용 된 키워드와 함께, 범죄에 사용되는 연락 수단을 포함하는 검색을 해 보았습니다. Google 검색 시, + 기호는 해당 단어를 포함한다는 뜻으로, 모든 검색 문장 뒤에 +카톡을 붙여 보았습니다.

검색	검색 결과	유해성 (1 페이지 기준)
대마 삽니다 +카톡	약 25,600개	매우 높음
엑스터시 삽니다 +카톡	약 6,620개	매우 높음
게임머니 환전 +카톡	약 2,900,000개	매우 높음
명의거래소 +카톡	약 560,000개	매우 낮음
섹스 파트너 +카톡	약 295,000개	매우 높음
뽕 삽니다 +카톡	약 145,000개	매우 높음
몰리 삽니다 +카톡	약 198,000개	매우 높음
머니상 +카톡	약 3,100,000개	매우 높음
장집 +카톡	약 13,800개	매우 높음
섹파 +카톡	약 186,000개	매우 높음

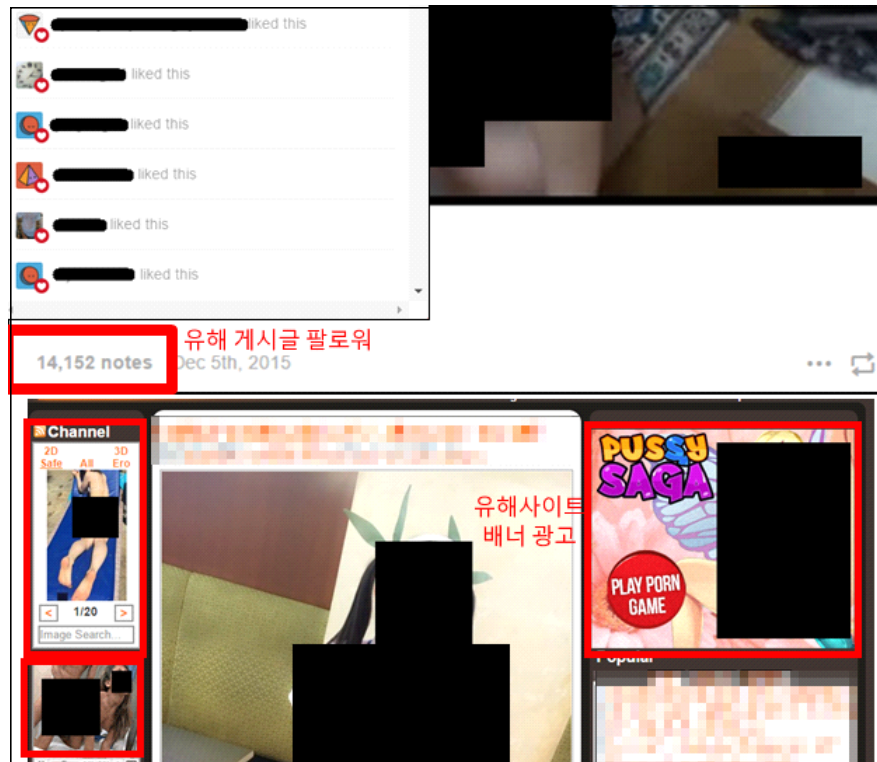
[표 2-3] 구글 검색 옵션 사용

+카톡을 붙인 것만으로도 일부 유형에서는 유해성이 극히 높아졌으며 (페이지 내 8건 이상), 범죄 조직들의 홍보 게시글로 검색 결과가 도배되었습니다. 특히 은어 뒤에 +카톡을 붙인 검색 결과는 모두 유해게시글이었을 뿐만 아니라, 카카오톡 연락처가 남았기 때문에, 증거자료 및 추적단서 수집에도 용의하였습니다.

2.3.4 유해사이트 내에서 탐색

직접 검색을 통해 유해사이트 혹은 광고/유해 게시글을 발견했다면, 해당 사이트 내의 측면 배너 혹은 팔로워 목록 등을 통해 유사한 유해 사이트를 방문할 수 있습니다. 아래 [그림 2-18]을 통해 알 수 있듯, 하나의 유해 게시글을 발굴 시, 그 아래 팔로워 대부분이 유해 게시글을 소지하고 있었으며, 유해사이트 속의 배너광고를 통할 경우, 더 다

양한 유해사이트를 방문할 수 있었습니다.



[그림 2-18] 유해 매체는 링크를 타고

이처럼 유해 미디어를 발견했다면 신고하고 끝내는 것이 아니라, 그 안에서 심층 모니터링을 진행하는 것이 새 게시글을 검색하는 것 보다 더 효율적이었으며, 음란물의 경우 링크를 타고 끝까지 파고들면 해당 유해 자료의 소스 URL 까지 차단시킬 수 있었습니다.

제 3 장 유해사이트 탐색 크롤러

3.1 유해사이트 탐색 크롤러 알고리즘

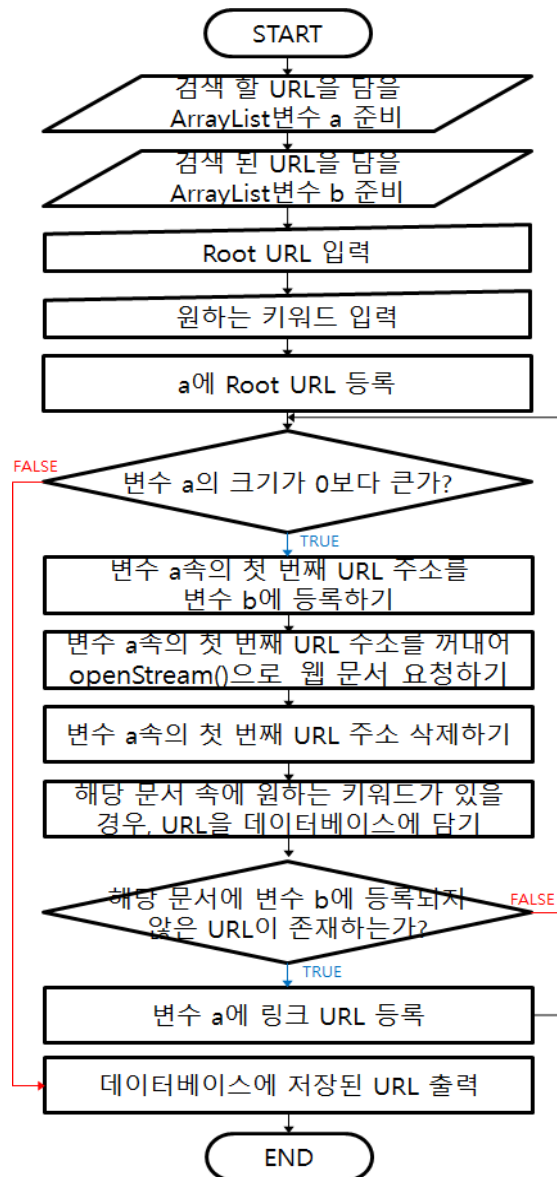
2장에서 연구한 내용을 토대로, 유해사이트들은 광고 혹은 직접적인 링크를 통해 이어져 있는 경우가 많기에 “향을 쫓 종이에선 향내가 나고, 생선을 쫓 종이에선 생선 내가 난다”는 옛말처럼 한 곳만 잡아도 그 주위로 연결된 수 십 수 백 개의 유해 사이트를 잡아내는 크롤러 알고리즘을 만들고자 하였습니다.

물론 초창기에는 이미 만들어져 있는 말테고(Maltego)프로그램을 이용하라는 동기들의 말을 들었지만, 말테고처럼 모든 인터넷상의 유해 정보를 모으게 될 경우, 원하지 않는 의학, 논문 등의 자료 또한 수집될 확률이 높았고, 궁극적으로 유료 프로그램이었기에 누리캡스 및 수사대원 분들에게 배포할 수 없는 상황이었습니다.

그래서 0부터 시작하여 99까지 뒤지는 대중적인 크롤링 방식 대신, 시작 URL을 지정하여 연결된 하이퍼링크를 따라 찾아 들어가는 무료 심층 탐색 크롤러를 만들기로 하였습니다.

순서	크롤러 동작 순서
1	시작 URL 지정
2	웹 문서 파싱
3	문서 속 하이퍼링크, iframe등 연결된 문서 수집
4	문서 내 유해 키워드가 있다면, 키워드 개수 및 해당 URL 수집
5	문서를 모두 검사했을 경우, 수집된 다른 문서에게 문서 내용 요청
6	수집 중단 명령이 있기 까지 순서 3으로 되돌아감
7	수집 중단 명령이 내려지면 분석을 위해 서버로 데이터 전송
8	클라이언트 프로그램 종료 및 분석 표 출력.

[표 3-1] 유해사이트 탐색 크롤러 동작 순서



[그림 3-1] 유해사이트 탐색 알고리즘

이외의 동작 고려사항으로는, 누리집스 회원과 수사대 분들을 돕기 위한 프로그램으로써, ‘편의성’과 ‘안전성’을 중시해야했습니다. 때문에 크롤러를 사용하면서 바이러스에 노출되지 않아야 했고, 역추적 가능성을 고려하여 프록시 기능이 필요하였습니다.

안전성 문제를 해결하기 위해 HTTP 프로토콜만 사용하고, JavaScript나 Flash 콘텐츠는 실행시키지 않도록 하여 악성 스크립트를 방지하고 속도를 높이는 방향을 선택하였습니다. 물론 이 때문에 일부 사이트 내에서는 페이지 탐색이 원활하지 않을 수도 있으나, 탐색 대상이 유해사이트이기에, 탐색능력을 양보하고 안전성을 높이는 방법을 선택하였습니다.



[그림 3-2] 필자가 만들 크롤러는 랜더링을 하지 않는 방식을 채택하였습니다

3.2 버전별 실험 및 평가

3.2.1 실험 환경

실험 순서는 업데이트 순으로 진행할 것이며, NuriCHAIN beta 1버전 이하의 실험은 [표 3-2]의 컴퓨터 환경에서 이루어 졌으며, NuriCHAIN beta 2버전의 실험은 [표 3-3]의 컴퓨터 환경에서 이루어 졌으며, NuriCHAIN alpha 1버전의 실험은 [표 3-4]의 컴퓨터 환경에서 이루어 졌습니다.

항목	값
OS	Windows 8 Enterprise (x64)
CPU	i7 - 4790K
Internet	1.0 Gbps
JAVA	JDK - 1.8.0_91 (x64)
메모리 할당량	128MB ~ 1GB
Root URL	http://ytteyteyte.tumblr.com/ (실제 유해 SNS 계정)
키워드	sex, porn, pussy, fuck, dick, young, rape, hentai
검색 시간	10분

[표 3-2] DragonEye, NuriCHAIN beta 1버전 실행 환경

항목	값
OS	Windows 10 Home (x64)
CPU	i7 - 6700HQ
Internet	1.0 Gbps
JAVA	JDK - 1.8.0_121 (x64)
메모리 할당량	128MB ~ 1GB
Root URL	7개 (실제 유해 SNS 계정)
키워드	65개의 한글(60) 영문(5) 키워드
스레드	80개
검색 시간	10분 (프록시 미 사용)

[표 3-3] NuriCHAIN beta 2버전 실행 환경

항목	값
OS	Windows 10 Home (x64)
CPU	i7 - 6700HQ
RAM	16 GB
Internet	1.0 Gbps
JAVA	JDK - 1.8.0_121 (x64)
메모리 할당량	128MB ~ 1GB
Root URL	9개 (실제 유해 SNS 계정)
키워드	165개의 한글(155) 영문(10) 키워드
스레드	8
검색 시간	10분 (프록시 사용)

[표 3-4] NuriCHAIN alpha 1버전 실행 환경

또한 사진 속 화면을 그대로 캡처 할 경우, 유해 사이트 URL 목록이 모두 나오기 때문에, 이 문서를 읽을 독자 분들을 위해 Root URL로 <http://google.co.kr>을 잡고, google 이란 키워드를 찾는 형태로 캡처 하였습니다.

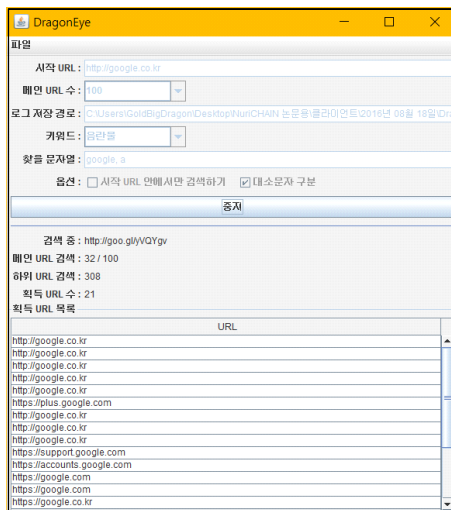
3.2.2 DragonEye

DragonEye는 NuriCHAIN이라는 이름을 짓기 전 필자의 이름을 본 따 만든 임시 이름으로, 가장 최초의 버전입니다.

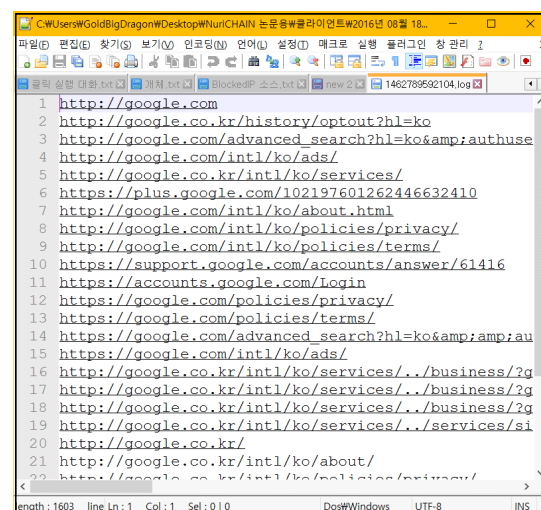
등록한 키워드가 웹 문서에 포함되어 있다면, 해당 URL을 수집하는 형태로, 가장 단순하지만 아래 [그림 3-3]처럼 찾았던 URL을 또 다시 찾아 버리는 버그가 있던 조잡한 버전이었습니다.

대상	값
총 검색 메인 URL 개수	575개
획득 URL 수 (중복 제외)	20개
실제 유해 사이트 개수	13개

[표 3-5] DragonEye 실행 결과



[그림 3-3] DragonEye 실행 화면



[그림 3-4] DragonEye 출력 로그

3.2.3 NuriCHAIN beta 1.0.0

유해사이트 탐색 크롤러 제작에 본격적으로 들어가기 전, 누리캡스(Nuri Cops)의 ‘누리’와 네트워크상의 유해한 광고들을 잡는다는 뜻의 Catch the Harmful Ad In Network의 약자 ‘CHAIN’을 합쳐 NuriCHAIN이라는 이름을 지어 주었습니다.

DragonEye의 중복 URL 수집 버그를 “List 변수를 생성하여 수집하는 URL마다 이전에 수집되었는지 비교하는 구문”을 작성함으로 써 수정하였으며, 획득 URL 뿐만 아니라 키워드 점유율 까지 수집 하여 보고서 출력시 도움이 되도록 업데이트 하였습니다.

패치 이후 [표 3-6]과 같이 검색 효율이 전체적으로 소폭 상승한 모습을 보였습니다.

대상	값
총 검색 메인 URL 개수	636개 (+61)
획득 URL 수 (중복 제외)	31개 (+11)
실제 유해 사이트 개수	14개 (+1)

[표 3-6] NuriCHAIN beta 1.0.0 실행 결과

3.2.4 NuriCHAIN beta 1.0.1

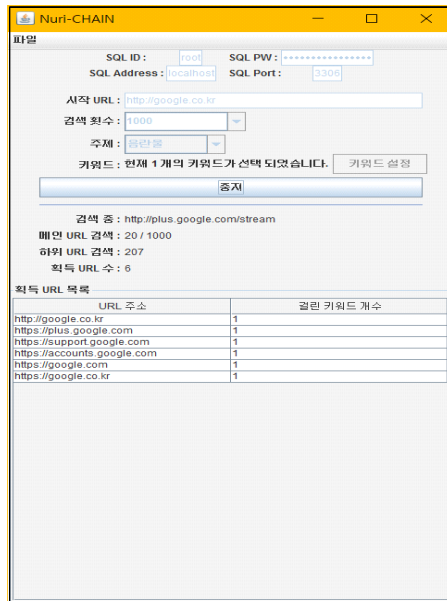
편의성 향상을 위해 키워드 설정 창을 추가시켜 프로그램이 실행중인 상태에서 키워드를 수정할 수 있게 하였으며, 프로그램 실행 시, DB가 구축되어 있지 않을 경우, 자동으로 생성되게 하였습니다.

탐색 효율 향상을 위해 “상위 URL 주소가 동일할 경우 수집하지 않도록”하여, 검색 횟수는 30건 가량 줄었지만 실제 유해 사이트 개수는 4건 증가하였습니다.

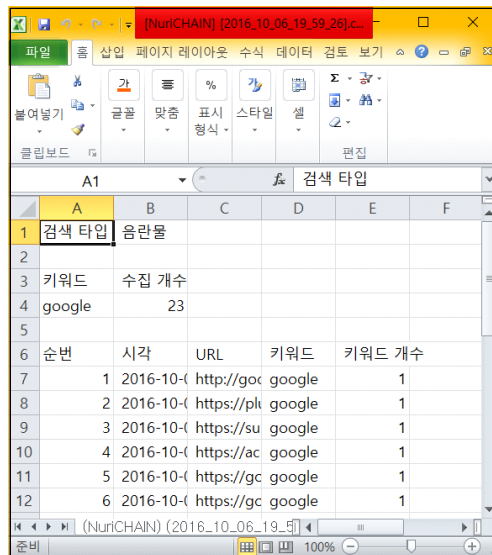
대상	값
총 검색 메인 URL 개수	597개 (-39)
획득 URL 수 (중복 제외)	30개 (-1)
실제 유해 사이트 개수	18개 (+4)

[표 3-7] NuriCHAIN beta 1.0.1 실행 결과

그 외에도 로그 추출 방식을 일반 텍스트(.txt)에서 엑셀 문서(.csv)형태로 변환하여 보고서 형태를 잡아 나가기 시작하였습니다.



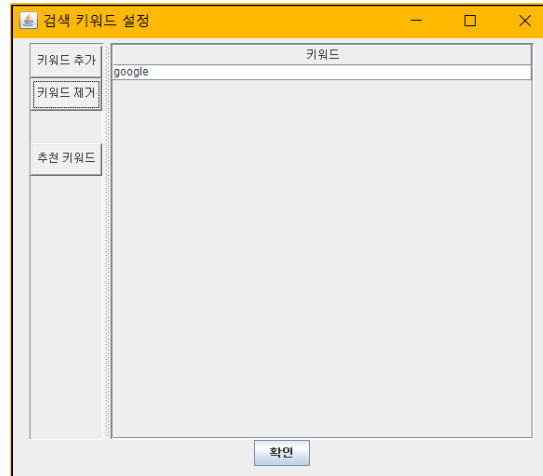
[그림 3-5] 키워드 버튼이 생긴 모습



[그림 3-6] CSV 형식 로그 출력 모습

3.2.5 NuriCHAIN beta 1.0.2

배포 이후, 어떤 키워드를 가지고 크롤링을 시작해야할지 모를 신입 누리캡스 회원의 사용을 고려하여 데이터 수집 서버를 개설하였습니다. 데이터 수집 서버는 NuriCHAIN을 통해 얻어낸 키워드 데이터를 수집하였으며, 수집된 데이터 중, 가장 많은 URL을 낚은 상위 10개의 키워드를 선별하여 키워드 목록에 추가 시켜 주는 [추천 키워드] 버튼을 추가하였습니다.



[그림 3-7] 추천 키워드 요청버튼 추가

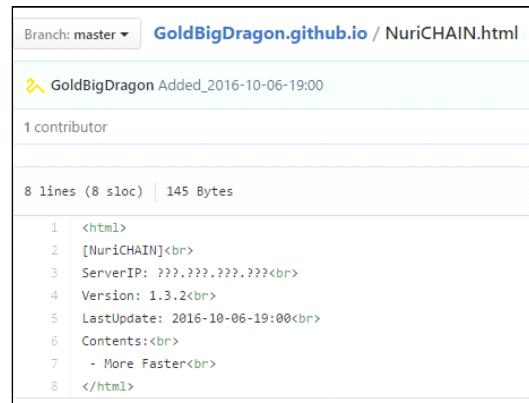
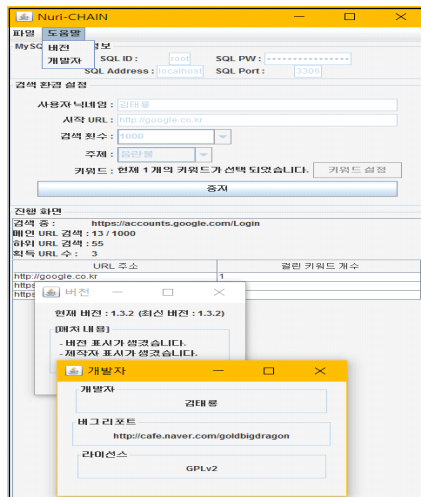
3.2.6 NuriCHAIN beta 1.0.3

검색 속도를 향상시키기 위해 “크롤링 클래스 리팩토링” 및 “키워드 연산 변수 타입을 String에서 Enum값 계산으로 변경”하여 연산 속도를 향상시켰습니다. 또한 도움말 메뉴 목록에 버전 및 개발자 정보를 추가하여 현재 버전 확인과 버그 리포트를 할 수 있게 만들어 배포 이후 피드백을 받을 수 있도록 하였습니다.

속도는 소폭 상승하였지만, 연구가 진행되는 동안 일부 유해사이트가 차단되면서 실제 유해사이트 개수가 소폭 줄어든 결과를 볼 수 있었습니다.

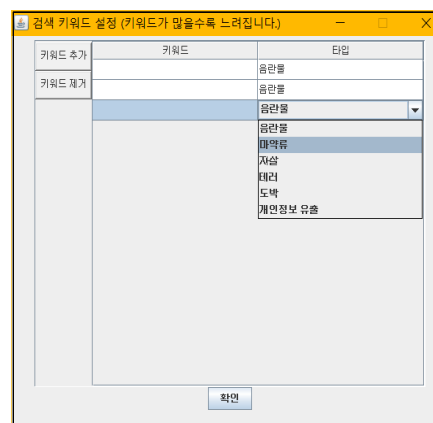
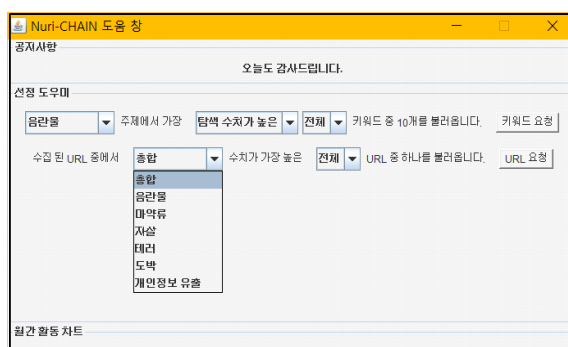
대상	값
총 검색 메인 URL 개수	663개 (+66)
획득 URL 수 (중복 제외)	30개
실제 유해 사이트 개수	15개 (-3)

[표 3-8] NuriCHAIN beta 1.0.3 실행 결과



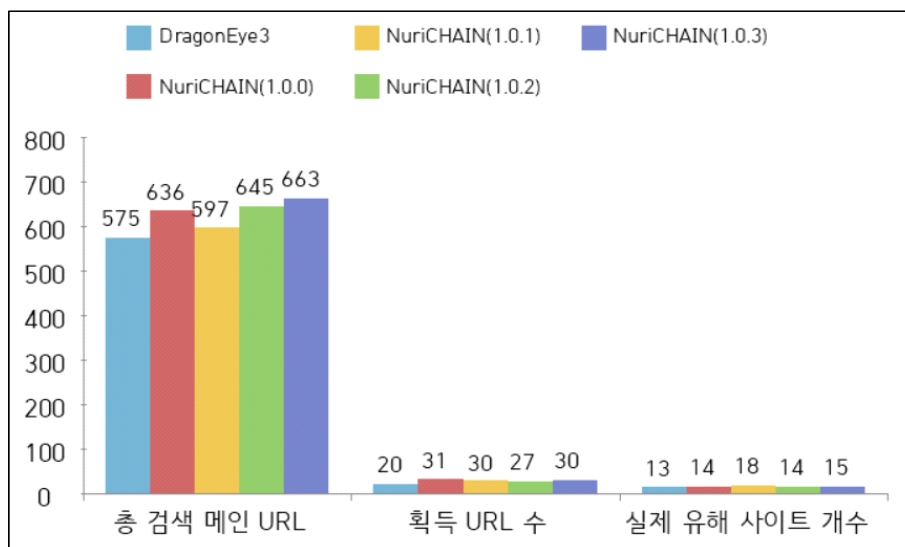
3.2.7 NuriCHAIN beta 1.0.4

키워드 별 타입을 지정 할 수 있도록 하여 한 번에 여러 타입의 유해 사이트를 수집할 수 있게 하였습니다. 또한 어떤 URL을 시작점으로 잡아 크롤링을 해야 할지 모를 신입 누리caps 회원의 사용을 고려하여 “수집된 URL 중 가장 많이 검색 된 URL 중 하나”를 지정해 주는 기능을 추가하였습니다.



3.2.8 NuriCHAIN beta 1 조건부 실험

[그림 3-11]은 DragonEye부터 NuriCHAIN 1.0.4버전까지의 성능 분석 차트로, 업데이트가 검색 속도 상승에만 영향을 줄 뿐, 유해 사이트 정확도에는 기여를 거의 하지 못하는 모습을 볼 수 있습니다.



[그림 3-12] NuriCHAIN beta 1 성능 분석 차트

이는 업데이트 때 마다 속도가 아무리 증가한다 하여도 정확도의 차이가 없다는 말로, 정확도를 올리기 위해서는 지금껏 실험 도중 변경하지 않은 환경요소인 ‘키워드’를 변경해 봐야한다는 결론이 나옵니다.

특히 앞전 실험에 사용된 sex 키워드는 의학계 용어 및 성별을 묻는 기사나 논문에도 자주 등장하는 단어이며, young은 말할 것도 없고, fuck은 한국어 ‘제길!’이란 뜻이 있기 때문에 각종 블로그나 트위터에 자주 올라오는 단어였습니다. 그래서 해당 단어가 하나라도 섞여 있던 가치 없는 URL이 많이 걸려들었으며, 이들이 정확성을 떨어뜨린 원인이 되었습니다. 이 외에도 여러 원인이 있으리라 생각되어 다양한 실험을 해 보기로 하였습니다.

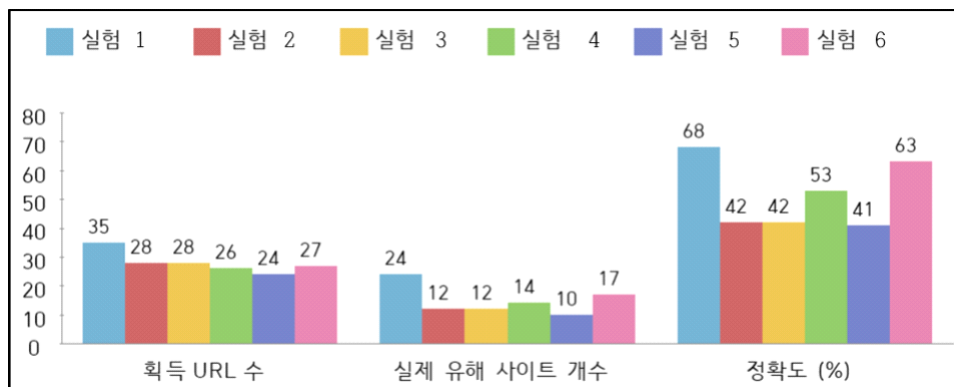
때문에 실험 환경을 조금 바꿔 10분간 NuriCHAIN 1.0.4버전에서 실제 유해 SNS 계정을 대상으로 크롤링을 진행하기로 하였습니다.

항목	값
OS	Windows 8 Enterprise (x64)
CPU	i7 - 4790K
RAM	28 GB
Internet	1.0 Gbps
JAVA	JDK - 1.8.0_91 (x64)
메모리 할당량	128MB ~ 15GB
Root URL	http://yteyteyte.tumblr.com/ (실제 유해 SNS 계정)
NuriCHAIN	1.0.4 버전
총 검색 메인 URL	1000개가 될 때 까지 검색

[표 3-9] 조건부 실험 환경

실험 번호	실험 내용
1	가장 많은 URL을 수집한 상위 10개의 키워드 사용
2	1의 조건 그대로 가장 많이 수집된 URL을 Root URL로 설정
3	유해사이트를 가장 많이 잡은 상위 10개의 키워드 사용
4	가장 많은 키워드가 걸린 URL을 Root URL로 설정
5	3의 조건을 유지한 채로 sex 키워드 제거
6	4의 조건을 유지한 채로 sex 키워드 제거

[표 3-10] 실험 내용



[그림 3-13] 조건부 실험 결과 차트

실험 1을 통해 "URL 수집률이 높은 키워드 교체"만으로 정확도가 가장 높았던 DragonEye3보다 정확도가 3%가량 증가한 모습을 확인 할

수 있었습니다.

실험 2와 3을 통해 “가장 많이 언급 된 URL을 RootURL로” 주는 행위나 단순히 “실제 유해사이트를 많이 잡은 키워드를 선정”하는 것은 정확도와 수집률에 큰 영향을 미치지 않는 것을 알 수 있었습니다.

실험 4를 통해 RootURL을 “가장 많은 키워드가 걸렸던 URL”로 설정 할 경우, 정확도가 9% 상승한 모습을 볼 수 있었습니다.

마지막으로 키워드 중, 가장 쓰레기 데이터를 많이 모았던 sex 키워드를 삭제 했더니, 실험 3과 동일 조건에서는 정확도가 1%줄었지만, 실험 4와 동일 조건에서는 정확도가 10%나 상승한 모습을 볼 수 있습니다. 또한 실험 3보다 정확도가 줄은 이유는, 수집 된 데이터, 즉 모집단이 줄었기에 전체적으로 본다면 정확도를 상승 시켜 주는 요인이 된 것을 알 수 있었습니다.

실험을 정리하자면 ‘가장 많이 수집 된 키워드’에 ‘가장 많은 키워드가 걸린 URL’을 사용하고, ‘쓰레기 키워드를 제거’한 상태에서 가장 높은 정확도를 낼 수 있었습니다.

이처럼 유해사이트를 엄청나게 많이 잡아 본 베테랑 사용자의 경우, 좋은 키워드와 URL을 주입 하여 높은 정확도가 나오고, 초보자의 경우 낮은 정확도가 나오기 때문에, 사용자가 제시 한 키워드만으로 크롤링 하는 프로그램은 당연히 정확도가 떨어질 수밖에 없으므로, 좋은 키워드를 제공 하는 기능이 필요하다는 뜻도 됩니다. 이는 클라이언트 프로그램에서만 개선책을 찾을 것이 아니라, 베테랑 사용자들의 도움으로 신선하고 품질 좋은 키워드를 제공 해 주는 서버 또한 다루어야 한다는 결론을 내릴 수 있습니다.

3.2.9 NuriCHAIN beta 2.0.0

2017년 초, beta 1 버전을 가지고 울산지방 경찰청에 직접 방문 시연을 하였습니다. 하지만 디자인이 너무 개발자 스타일이라는 말과 함께, 탐색 속도 면에서도 필자의 모니터링 속도보다 떨어져서 “프로그램을 쓰느니 너를 복제해서 쓰는게 낫겠다.”는 답변과, NuriCHAIN이 불순한 의도로 사용될 경우, 역으로 유해사이트수집기가 되어버릴 수 있다

는 충고를 들었습니다. 이에 누리캡스 회원이 아니라면 사용하지 못하도록 로그인 API를 제공해 줄 수 있느냐 물었지만, 해당 부분은 울산지방 경찰청 관할이 아니라 불가능 하다는 답변을 들었습니다.

집에 돌아와서 울산지방 경찰청에서 얻은 피드백과, NuriCHAIN beta 1 버전에서 극복하지 못한 문제를 정리 해 보니, 디자인과 속도, 안전성 3가지가 가장 큰 문제였습니다.

우선 단일 스레드를 사용하기 때문에, 웹 문서 한 페이지를 모두 훑고 가야만 다음 페이지로 넘어가서 필자의 모니터링 속도보다 느렸고, 최적화 동작을 사용하지 않아서 메모리상에 유해사이트 URL을 너무 많이 수집하게 될 경우, 메모리부족으로 컴퓨터가 느려지거나, 프로그램이 중단되었으며, 프록시를 사용하지 않아 Twitter 및 Tumblr 탐색 도중 DDOS 공격으로 오해받아 10분 내에 IP가 차단되는 등, 다양한 문제점이 있었습니다.

NuriCHAIN beta 1	NuriCHAIN beta 2
기본 제공 Windows 화면 GUI	그림판으로 그린 GUI
단일 스레드 사용	다중 스레드 사용
최적화 동작 사용하지 않음	최적화 스레드 사용
프록시 사용하지 않음	프록시 사용
한글 키워드 검색 불가	한글 키워드 검색 가능
검색 환경 저장 불가	검색 환경 저장/불러오기 가능
서버 관리 페이지 없음	서버 관리 페이지 개설
1시간 이내 메모리 초과로 작동 중지	최소 2일 이상 정상 가동
페이지 내의 URL만 수집	SNS 타입에 따라 추가 URL 수집
위키, 의학 사이트 URL 수집	필터링 기능 추가

[표 3-11] NuriCHAIN beta 2의 업데이트 내용

때문에 개발자스러운 GUI를 그림판으로 그림을 그려 기본 디자인을 변경하여 딱딱함을 없앴습니다.

단일 스레드의 경우, 스레드 추가 코드를 기입하여 이용자가 원하는 성능을 끌어 낼 수 있도록 스레드 개수 기입란을 추가하였습니다.

URL이 계속 메모리에 쌓여 결국 터져버리는 문제는 URL 개수가 일

정 개수 이상 넘어갈 경우, 자동으로 파일로 저장해 두거나, 소거하여 3만개 이상의 URL을 메모리에 쌓아두지 않도록 최적화 시키는 스레드를 추가하여 메모리 부족 현상을 해결하였습니다.

프록시를 사용하지 못하여 차단당하는 일을 없애기 위해 프로그램 시작 전, 프록시 주소들을 적어 둘 경우, 해당 주소로 프록시를 이용하도록 설정하여 차단을 우회하도록 하였습니다.

한글 키워드는 검색 불가하던 버그를 UTF-8 인코딩 변환 코드를 작성함으로써 해결하여 한글 키워드 또한 검색이 가능하게 하였습니다.

검색 시작 전, 검색 키워드를 일일이 입력해야 하고, URL 또한 입력해야하는 번거로움과, 누리캡스 혹은 수사대간 키워드 및 URL을 공유할 때를 생각하여 현재 설정된 모든 검색환경을 저장 및 불러오기 할 수 있는 편의기능을 추가하였습니다.



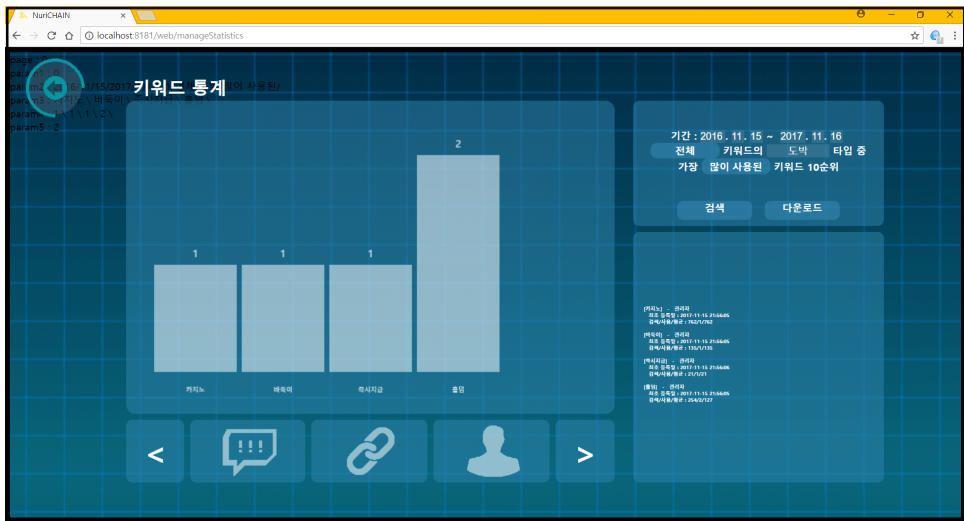
[그림 3-14] NuriCHAIN beta 2의 메인 화면

SNS(Tumblr, Twitter 등)를 모니터링 할 경우, 이용자의 팔로잉 계

정 및 좋아요 표시한 게시물까지 SNS 타입에 따라 추가 URL을 수집하도록 변경하여 모니터링 연계성을 높였습니다.

URL에 google 하위 문서 및 위키 등이 도메인일 경우, 수집하지 않도록 설정하여 탐색 속도 향상 및 정확도를 향상시켰습니다.

또한 수집된 키워드, URL을 관리 및 통계를 낼 수 있는 웹 페이지를 개설하여 신입 누리caps 및 수사관이 키워드/URL을 확인할 수 있도록 하였습니다.



[그림 3-15] 웹 페이지에서 제공하는 통계 그래프

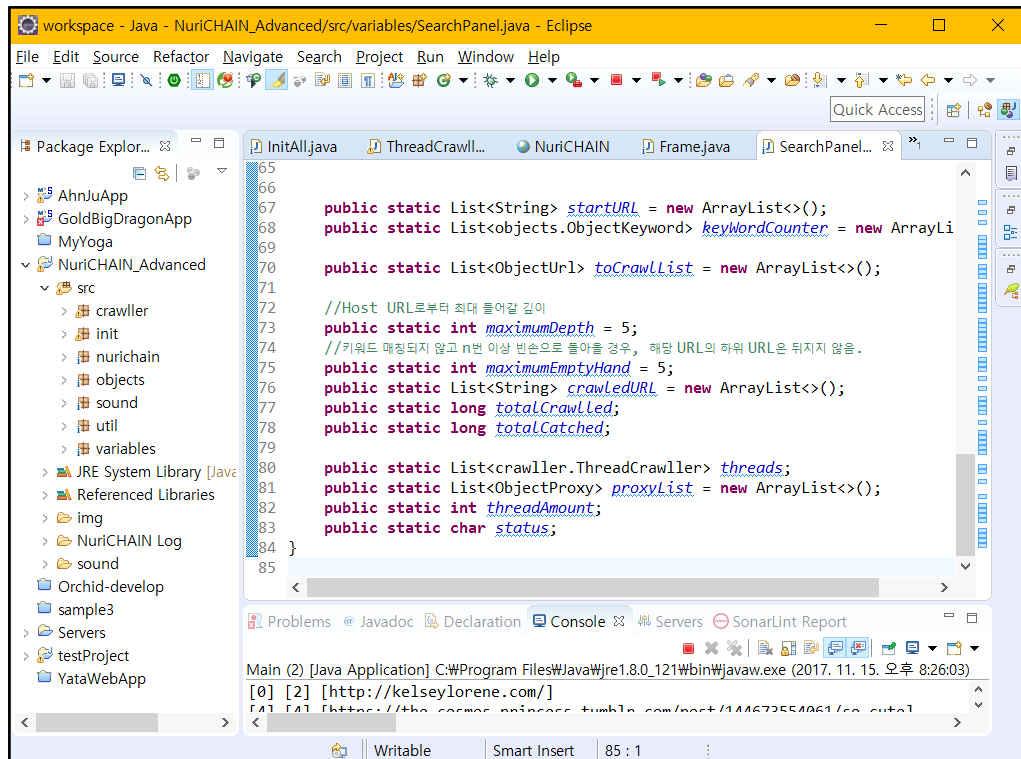
웹 페이지 개설 이후 프록시를 사용하지 않은 상태로 10분간만 모니터링을 실행 해 보았습니다.

대상	값
획득 URL 수 (중복 제외)	1303개
실제 유해 사이트 개수	1079개
정확도	82%

[표 3-12] NuriCHAIN beta 2의 모니터링 결과

단순 수치만 보더라도 이전 버전에 비해 상당한 발전을 이루었으며, 획득 URL 수와 실제 유해사이트의 비례 관계, 즉 정확도의 개선은 상당히 좋은 소식인데, 이는 “동일 도메인을 가진 URL의 경우, 어디까지 탐색 할 것인가”를 정하는 ‘최대 깊이’ 변수와, “일정 횟수 이상 키워드에 매칭되지 않은 사이트를 돌았던 메인 도메인의 경우, 자동으로 탐색 리스트에서 삭제”하는 ‘빈손 알고리즘’ 넣었기 때문입니다.

또한 검색에 사용된 시작 URL과 키워드는 DragonEye 버전부터 쌓아온 데이터 통계로 얻은 것으로, 통계 및 집계 표를 나타내어 주는 웹 페이지가 한 몫 하였습니다.



[그림 3-16] maximumDepth, maximumEmptyHand로 성능을 향상시켰다!

NuriCHAIN beta 2 시연 영상 : <https://youtu.be/27su8FK32BE>

3.2.10 NuriCHAIN alpha 1.0.0

NuriCHAIN beta 2버전 완성 이후, 더욱이 누리캡스 회원들에게 배포하고 싶은 욕구가 불타올라서 어떻게든 누리캡스 API를 얻고 싶었습니다. 하지만 누리캡스 담당 기관에 연락할 만한 뽕족한 수가 없었기에, 한 가지 묘안을 생각 하였습니다.

누리캡스는 1년에 단 한 번, 신고 활동이 우수한 회원을 뽑아 ‘베스트 누리캡스’ 시상식을 열고, 시상식 때는 경찰청장 및 사이버 수사대와 누리캡스 담당 경찰 분들이 오셔서 순위권 내의 누리캡스 회원들과 식사를 함께하는 것이었습니다. 따라서 2018년 1년간 최선을 다해 순위권 내에 든 다음, 담당 경찰관분께 직접 NuriCHAIN 소개 및 누리캡스 API 개방 여부를 물어 보는 계획을 세웠습니다.

물론 쉽지 않았습니니다. 필자가 NuriCHAIN을 개발 한다는 소식을 들은 다른 경찰청 내에서 필자가 매크로를 이용하여 1등을 한다는 유언비어를 퍼뜨리고, 해명글을 요구하는 등, 상하 좌우로 압박을 해 왔습니다. 하지만 필자는 지금껏 NuriCHAIN을 통해 수집했던 유해사이트는 누리캡스 점수로 환산하지 않았음은 물론, 자동으로 신고한 적이 없었기에 당당히 증거자료 제출과 신고에 임하였습니다. (증거 자료로는 8, 9, 10월 동안 신고하는 모습이 담긴 동영상 자료를 제출하였습니다. https://mega.nz/#F!hs8hSCYR!8HXJqX1YZo7vS_nPfgxGxA)



[그림 3-17] 베스트 누리캡스 상장 및 트로피

이런 저런 사정이 많았지만, 결국 1년간 약 4만 7천 건을 신고하여 베스트 누리캡스에 선정되었으며, 서울 본청, 수사대 분들이 보시는 앞에서 누리캡스 모니터링 팀에 대한 발표까지 하게 되었습니다. (간담회 자료 : <https://cafe.naver.com/goldbigdragon/81569>) 하지만 안타깝게도 역시나 API 요청은 거절되었습니다.

이후 NuriCHAIN 개발에 흥미를 잃어 이래저래 시간만 때우다가 2018년 12월, 게임물 관리 위원회로부터 불법 게임물 신고 감사장을 수여받고, 회식 자리에서 부산지방 경찰청 누리캡스 회원 분들을 만나게 되었습니다. 서로 간 격려가 오가는 와중, 크롤러 이야기가 나왔으며, 부산지방 누리캡스 회원님께서서는 웹 상 JavaScript로 돌아가는 크롤러를 만들어 자동 신고를 해 보셨다고 하셨습니다.

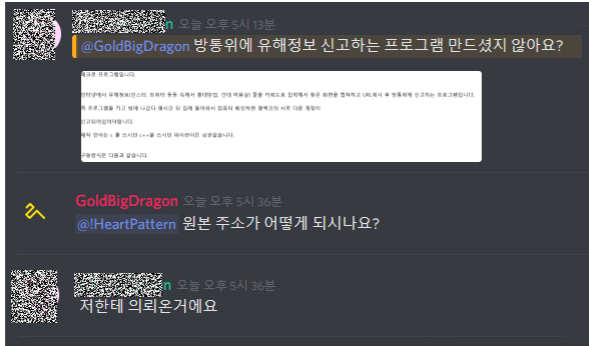
이에 JavaScript로 돌아가는 크롤러에 대해 동작 알고리즘을 생각하던 도중, 누리캡스 측에서 API를 주지 않아도, 누리캡스 회원인지 확인하는 방법이 떠올랐습니다. 바로, 크롤러 동작 원리를 조금만 비틀어서 누리캡스 홈페이지에 직접 로그인 시킨 다음, 누리캡스 홈페이지에서 주는 닉네임과 아이디를 가져오는 방식이었습니다.



[그림 3-18] 누리캡스 회원 인증 방식

API를 사용하지 않은 우회방법이지만, 사용자가 실제로 누리캡스에 로그인 하는 방식과 완벽히 일치한 구조이기 때문에, 경찰청 로그에도 접속 기록이 남고, 사용자도 NuriCHAIN 회원가입을 하지 않아도 되어 양측 모두 손해 보지 않는 방법이었습니다.

로그인 기능을 구현하고 나니, 때마침 사건이 터졌습니다. 지인 개발자로부터 누리캡스 자동 신고 매크로 제작 관련 외주가 들어오는 바람에, 이대로 19년도 누리캡스 신고대회가 시작될 경우 공정성을 해칠 것이 분명 해 보였습니다.



매크로 프로그램입니다.

인터넷에서 유해정보(인스타, 트위터 등등 속에서 흥대맛집, 건대 미용실) 들을 키워드로 입력해서 찾은 화면을 캡처하고 URL복사 후 방통위에 신고하는 프로그램입니다.

즉 프로그램을 키고 밖에 나갔다 몇시간 뒤 집에 들어와서 컴퓨터 확인하면 몇백건의 서로 다른 계정이 신고되어있어야합니다.

제작 언어는 c 를 쓰시던 c++을 쓰시던 파이썬이든 상관없습니다.

구동방식은 다음과 같습니다.

....

네 그렇게해도 괜찮습니다!! 사이트이름은 <https://www.nuricops.org/index.do> 여기에 신고하는거예요!

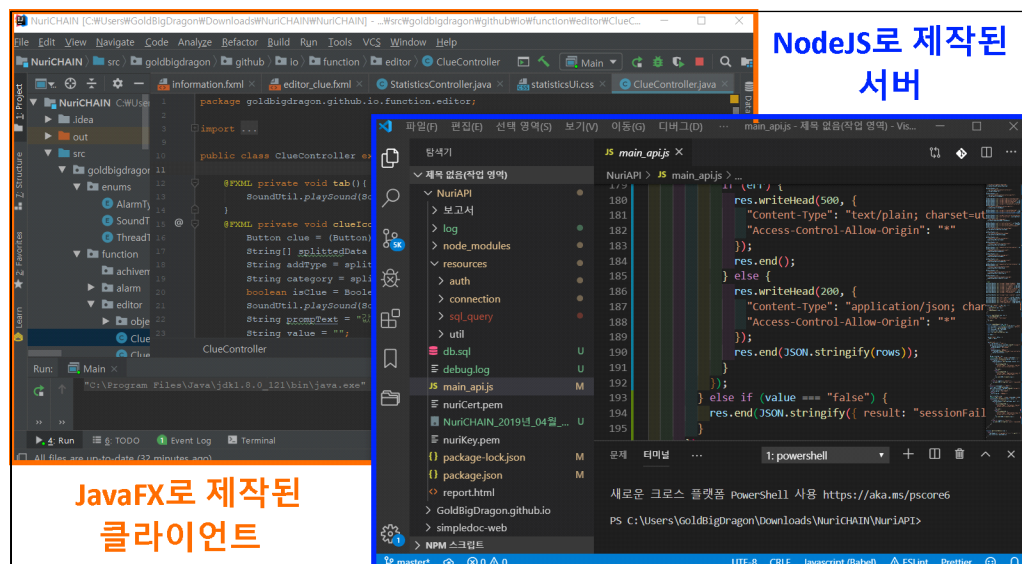
[그림 3-19] 지인 개발자에게 온 의뢰서 중 일부

안 그래도 몇 년 전부터 각종 지방청 아래 누리캡스 회원들이 개인

메일을 통해 NuriCHAIN 배포 관련 연락을 먼저 취해 왔었습니다만 당시 실험 단계였기도 하였고, 특정 지방청에만 먼저 배포하는 것은 공정성을 해치기 때문에 선불리 배포하지 못하였는데, 덕분에 NuriCHAIN을 눈치 안보고 배포할 수 있는 기회가 생긴 것입니다.

때문에 베스트 누리캡스 점수가 매겨지기 시작하는 4월 1일에 배포하는 것을 목표로 NuriCHAIN을 급히 제작해 나갔습니다.

NuriCHAIN beta 2 버전을 그대로 배포하자니 성에 차지 않아, 2018년 당시 IPP 장기 현장실습 프로그램을 통해 회사에서 사용하던 JavaFX를 이용하여 새로운 NuriCHAIN을 만들어 나가기 시작하였습니다. 서버 또한 SpringFramework 대신 NodeJS를 사용하여 가볍고 쾌적한 서버환경을 구성할 수 있었습니다. (IPP 장기 현장실습 : <https://cafe.naver.com/golbigdragon/84346>)



[그림 3-20] 클라이언트와 서버 개발 환경

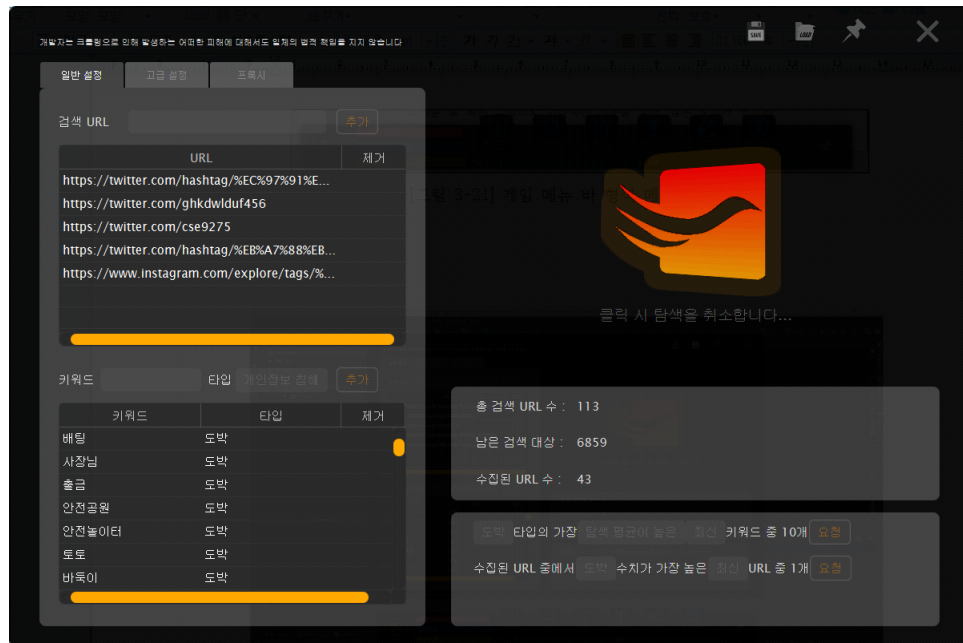
NuriCHAIN beta 2 버전이 성능 면에서 향상을 이루었기에, 배포용 버전은 ‘편리성’에 초점을 맞추기로 하였습니다. 대부분 대학생일 누리캡스 회원들을 위해 메인 GUI 및 각종 GUI를 게임과 비슷한 환경으로

맞추었으며, 효과음을 더하였습니다.



[그림 3-21] 게임 메뉴 바 형식 메인 메뉴 GUI

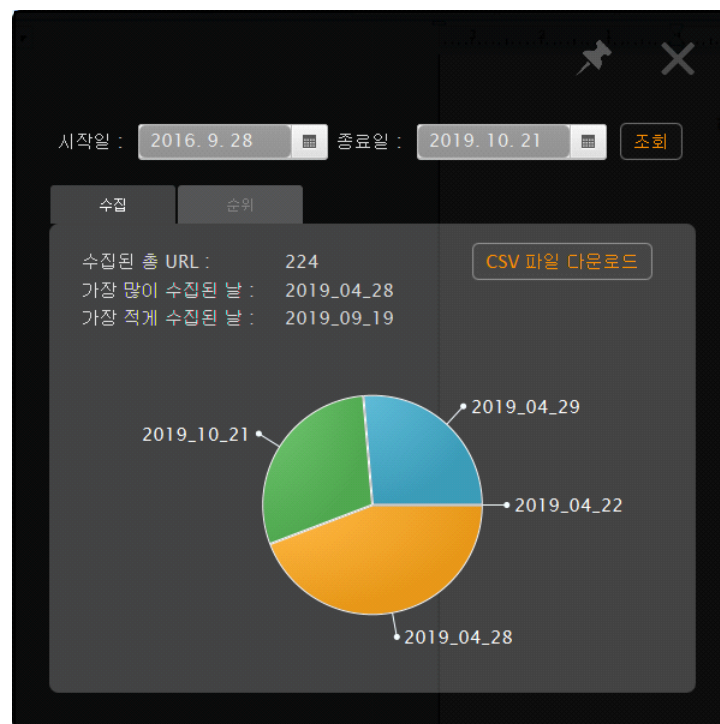
검색 알고리즘에는 변화가 없으나, 검색 GUI 디자인을 검은색으로 하여, 어두운 밤에도 눈이 부시지 않게 탐색할 수 있도록 하였습니다.



[그림 3-22] 탐색 GUI

탐색 이후에는 탐색 데이터를 CSV 파일로 출력 해 주며, 통계 메뉴에 들어가서 특정 기간 동안의 검색 통계를 확인/출력 해 줄 수 있도록 하였습니다. 통계의 경우, Pie Chart를 통해 시각적으로 유해사이트를 가장 많이 수집한 날을 알려주며, [순위]탭에서 유해 키워드가 가장 많이 걸린 URL 20순위 및 가장 많은 자식 URL을

가진 URL의 순위를 확인할 수 있게 하여, 누리캡스 및 수사관 분들을 편의성을 향상 시켰습니다.



[그림 3-23] 통계 GUI

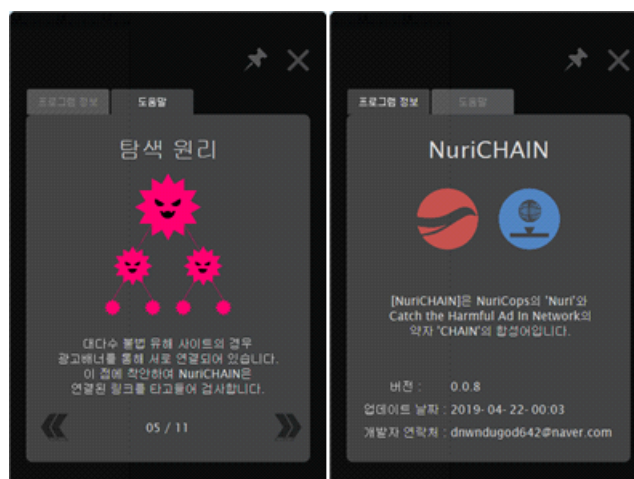
유해사이트를 신고할 경우, 신고 사유를 적는 칸이 있는데, 신입 누리캡스의 경우 작성 양식을 몰라서 ‘수사관님 언제나 수고하십니다.’ 등의 메시지를 입력하는 경우가 있었습니다. 때문에 신고 사유 작성을 도와줄 자동 서식 작성 기능도 추가하였습니다.

자동 서식 작성 기능의 또 다른 용도는, 수사관에게 도움이 되려면, 어떤 단서를 어떻게 제공해야하는지 알려주기 위함도 있었기에, 3년간 연구를 통해 추려낸 각종 단서 프리셋을 지원하였으며, 마우스를 올릴 경우, 어떠한 단서를 작성해야하는지 설명이 나오게 하였습니다.



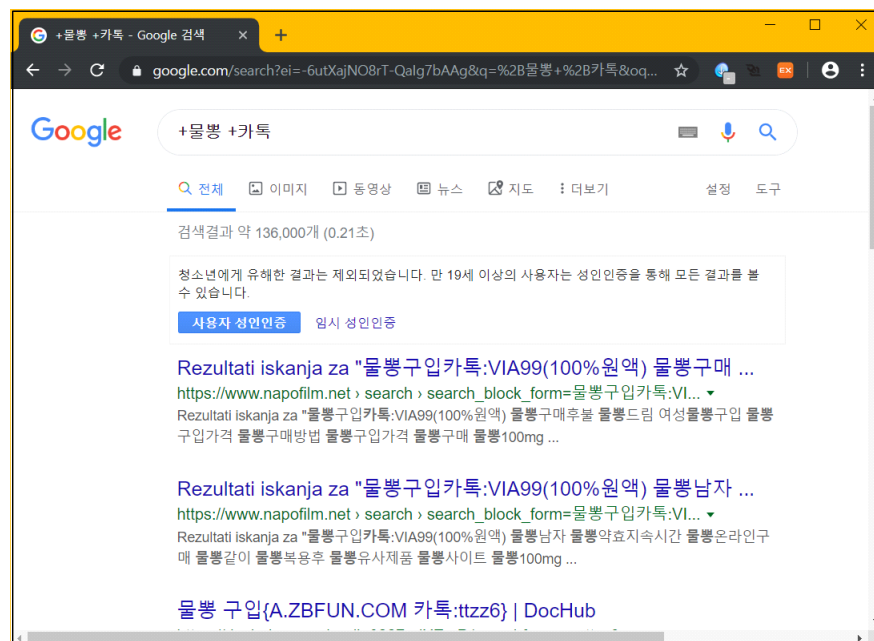
[그림 3-24] 자동 서식 생성 GUI

대부분의 GUI창 버튼 및 입력란에 마우스를 올릴 경우, 친절하 설명이 나오지만, 법적 고지 및 알고리즘 등의 도움말이 필요할 사용자를 위해 도움말 페이지도 잊지 않았습니다.



[그림 3-25] 도움말 GUI

대표적인 탐색, 통계, 서식 기능 외에도 아예 어디서 어떻게 찾아야 할지를 해매는 신입 누리캡스 회원을 위해 적중률이 높은 키워드를 서버로부터 받아온 다음, 이를 인터넷 브라우저에 입력하여 ‘임의의 유해 게시글을 발굴’하는 기능을 추가 해 두었습니다.



[그림 3-26] 임의의 유해 게시글을 발굴한 모습

Java에서 JavaFX로 넘어오면서 전체적인 코드 리팩토링을 통해 알고리즘을 최소화 시킨 덕에 테스트 가동 내내 메모리를 200 ~ 300MB 내에서만 점유 하였으며, NuriCHAIN beta 2 버전에서는 80 스레드를 사용하여 10분 만에 1303건 발견에 82%의 정확도를 보여주었으나, 현재 alpha 1 버전은 8 스레드를 사용하여 10분 만에 3098건을 발견 하였으며, 전체 URL 대상 정확도는 64%에 그쳤지만, 걸린 키워드가 6개 이상인 URL의 경우, 93.7%의 정확도를 보여주었습니다. 덕분에 수집된 URL 중, 키워드가 많이 걸린 URL일수록 유해사이트일 확률이 높다는 것을 확인할 수 있었으며, 그 기준은 6개 이상부터 가장 높은 정확도를 나타내었습니다.

[NuriCHAIN] [2019_10_21_1571666102104].csv - Excel

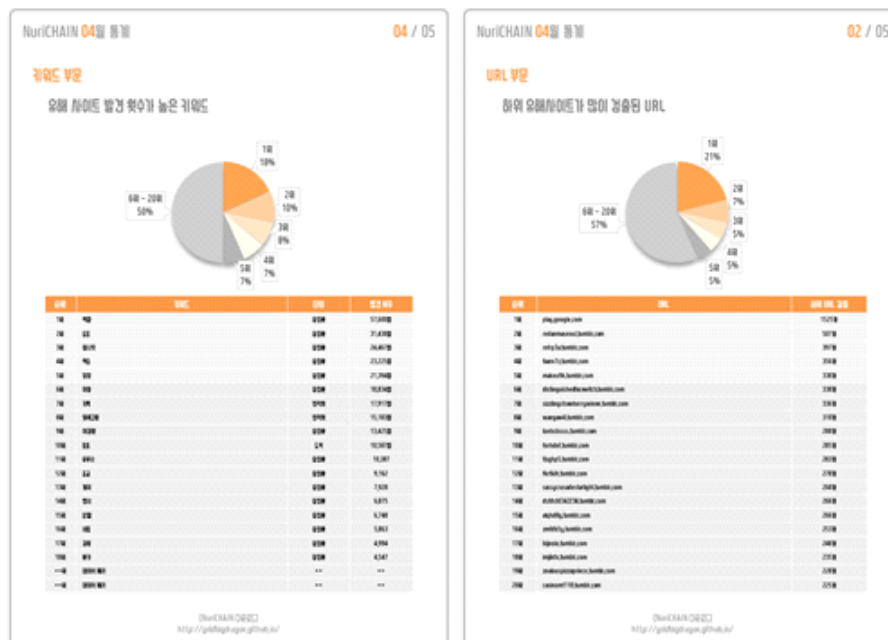
파일 홈 삽입 레이아웃 수식 데이터 검토 보기 도움말 어떤 작업을 원하시나요?

A1 검색자

	A	B	C	D	E	F
159						
160						
161	순번	시간	URL	키워드		키워드 개수
162	1915	2019-10-21 13:54:04	https://co	배팅, 카톡, 텔레그램, 아동, 섹스, 몰카, 섹파, 고딩, 유부녀, 보빨, 원나잇, 변녀, 오프, 여대생		14
163	668	2019-10-21 13:50:40	https://qc	토토, 카지노, 카톡, 텔레그램, 조교, 아동, 섹스, 네임드, 섹파, 유부녀, 오프, 여대생		12
164	1229	2019-10-21 13:52:13	https://lo	카톡, 강간, 초대남, 아동, 섹스, 초대녀, 쉬멜, 네토, 경병, 멜롬, 변녀, 오프		12
165	2192	2019-10-21 13:54:54	https://le	암캐, 조교, 조련, 초대남, 아동, 초대녀, 고딩, 보빨, 변녀, 사까시, 일탈, 오프		12
166	2	2019-10-21 12:41:23	https://tw	대마, 도리도리, 멜, 빙두, 뽕, 엑스터시, 작대기, 코카인, 크리스탈, 필로폰, 히로뽕		11
167	403	2019-10-21 13:50:00	https://wf	카톡, 조교, 아동, 섹스, 게이, 섹파, 유부녀, 원나잇, 일탈, 오프, 여대생		11
168	1617	2019-10-21 13:53:15	https://qv	강간, 수지, 암캐, 아동, 섹스, 몰카, 섹파, 고딩, 유부녀, 원나잇, 변녀		11
169	1957	2019-10-21 13:54:12	https://dh	사장님, 토토, 아동, 섹스, 후장, 뽕, 게이, 섹파, 원나잇, 일탈, 여대생		11
170	2215	2019-10-21 13:54:58	https://wr	포커, 아동, 섹스, 사다리, 게이, 섹파, 유부녀, 원나잇, 일탈, 오프, 여대생		11
171	97	2019-10-21 13:45:31	https://co	배팅, 출금, 토토, 바카라, 카지노, 섹파, 유부녀, 원나잇, 변녀, 오프		10
172	105	2019-10-21 13:45:43	https://ba	출금, 토토, 바카라, 카톡, 텔레그램, 수지, 아동, 사다리, 게이, 시디		10
173	258	2019-10-21 13:49:38	https://co	배팅, 출금, 토토, 바카라, 카지노, 섹파, 유부녀, 원나잇, 변녀, 오프		10

(NuriCHAIN) (2019_10_21_1571666)

[그림 3-27] 검색 결과를 CSV로 출력한 모습



[그림 3-28] 홈페이지에 게시 된 NuriCHAIN 4월 사용통계

이후 몇 가지 버그들을 고치느라 시간이 조금 흘렀지만, 2019년 4월 11일 누리캡스 홈페이지에 등록하여 전국 누리캡스 및 관계자

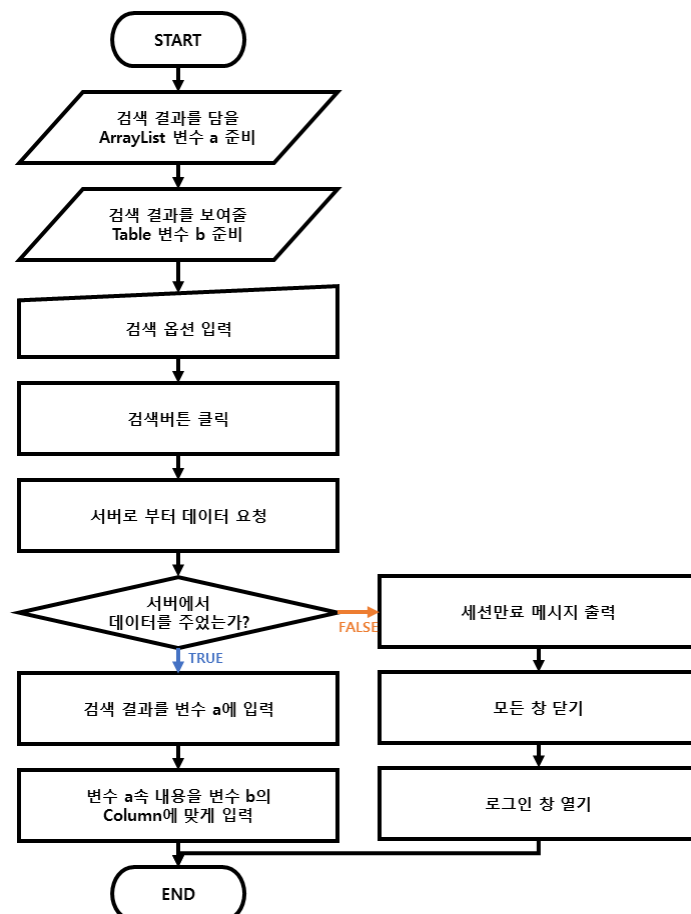
분들에게 NuriCHAIN을 배포하였습니다. 배포 이후 4월 한 달간 12명의 회원이 사용하였으며, 총 127,257건의 유해사이트를 탐색 해 주신 덕에 데이터가 많이 모였습니다. 이를 토대로 4월 통계 보고서를 작성하여 누리캡스 홈페이지에 게시하여 누리캡스 회원들에게 다양한 정보를 제공 하였습니다.

보고서를 몇 시간 결려 힘들게 직접 그림판으로 그리다 보니, 초창기에 기획 했던 ‘시각화’에 대한 욕망이 다시 꿈틀거리기 시작하였습니다. 당시 도박 및 불법 프로그램 관련 사이트를 뒤지면서, 범죄단서 추적 트리를 그려주는 프로그램을 구상하다가 Java 실력이 부족하여 구현을 미루었지만, 이제는 시각화에 특화된 JavaFX를 사용하는데다 전국 누리캡스 회원들뿐만 아니라 일부 경찰관 분들도 사용 해 보셨던 NuriCHAIN으로 더 순도 높은 빅 데이터를 모을 수 있었습니다. 환경이 좋아진 까닭도 크지만, 지금 까지 개발 해 온 NuriCHAIN은 누리캡스 회원들의 모니터링 및 신고용으로만 쓰였을 뿐, 수사에 실질적 도움이 될 만한 기능이 적었기에, 이제는 시각화를 하여 수사에 도움이 되는 프로그램을 만들 때라는 것을 본능적으로 느꼈습니다.

제 4 장 데이터 분석

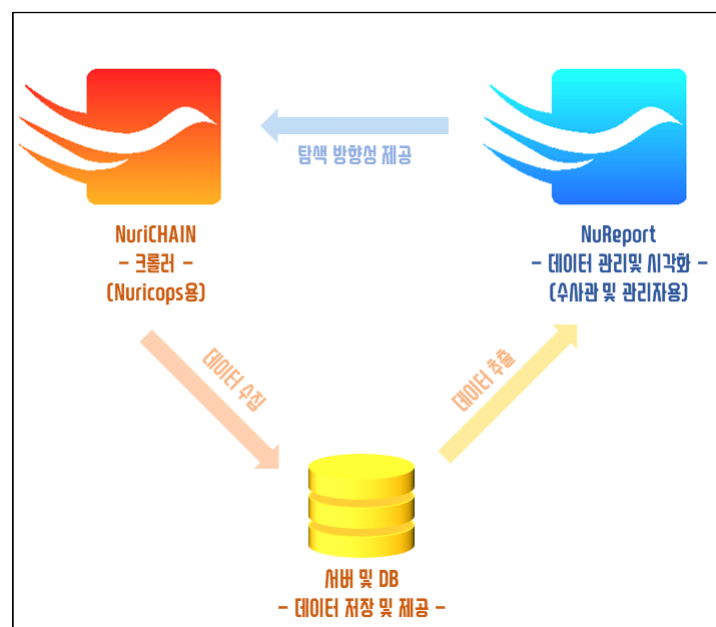
4.1 데이터 분석 알고리즘

NuriCHAIN을 통해 추적 단서, 키워드, URL 등의 데이터를 모은 다음, 분석 프로그램을 이용하여 해당 데이터를 관리 및 시각화 시키는 기본적인 알고리즘을 토대로, 누리캡스 회원이 아닐 경우, 서버로부터 데이터를 받을 수 없도록 암호화된 키 값을 부여하는 보안 알고리즘을 첨가하여 클라이언트-서버간 보안 통신이 가능하게 하였습니다.



[그림 4-1] 데이터 추출 알고리즘

NuriCHAIN을 통해 수집된 데이터를 분석하는 프로그램이기에, 누리캡스의 Nuri와 보고서의 Report를 합쳐 NuReport라는 이름을 지어 주었으며, NuriCHAIN을 통해 수집된 데이터는 서버를 통해 DB에 저장되고, DB에 저장된 데이터를 NuReport를 통해 불러와서 표 혹은 시각화 그래프로 나타내어 주는 방식입니다.



[그림 4-2] NuriCHAIN 크롤링 생태계

4.2 데이터 분석

4.2.1 CSV 파일 출력

검색 결과를 그대로 CSV 엑셀 파일 형태로 추출해 내는 기능입니다. 데이터양이 많을 경우, 20개당 1페이지로 나눠서 출력됩니다. 하지만 한 번에 많은 양의 데이터를 출력할 경우, 서버는 물론, 사용자 컴퓨터에도 부담이 가기 때문에, 서버로부터 데이터 요청 시 1000개 씩 끊어 불러오도록 하였으며, 도중에 다운로드를 중단 할 수 있도록 개

별 스프레드를 사용하도록 만들었습니다.

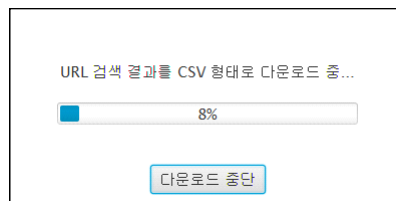
URL 관리 수집된 URL을 수정 및 삭제합니다.

2017. 10. 4 ~ 2019. 10. 23 기간 동안 수집된 모든 타입의 URL을 기본 오름차순으로 검색

최초 탐색일	최근 탐색일	URL	타입	상위 URL	탐색 횟수	최초 등록자	신뢰도	관리
2019-04-...	2019-04-...	http://...	연락처	eoduru...	0	김태룡	0	✖
2019-04-...	2019-04-...	http://...	연락처	eotjrd7...	0	김태룡	0	✖
2019-04-...	2019-04-...	http://...	연락처	http://...	0	김태룡	0	✖
2019-04-...	2019-04-...	http://...	음란물	http://...	0	김태룡	0	✖
2019-04-...	2019-04-...	http://...	음란물	http://...	0	김태룡	0	✖
2019-04-...	2019-04-...	http://...	연락처	http://...	0	김태룡	0	✖
2019-04-...	2019-04-...	http://...	음란물	fjldip.t...	0	김태룡	0	✖
2019-04-...	2019-04-...	http://...	음란물	http://...	0	김태룡	0	✖
2019-04-...	2019-04-...	http://...	도박	fkrtkih...	0	김태룡	0	✖
2019-04-...	2019-04-...	http://...	연락처	fkrtkih...	0	김태룡	0	✖
2019-04-...	2019-04-...	http://...	마약류	jojoget...	0	김태룡	0	✖
2019-04-...	2019-04-...	http://...	연락처	http://...	0	김태룡	0	✖

3 / 1756

[그림 4-3] 수집된 유해사이트검색 결과



[그림 4-4] 다운로드 화면

[NUReport] [URL] [1571841120322].csv 김태룡

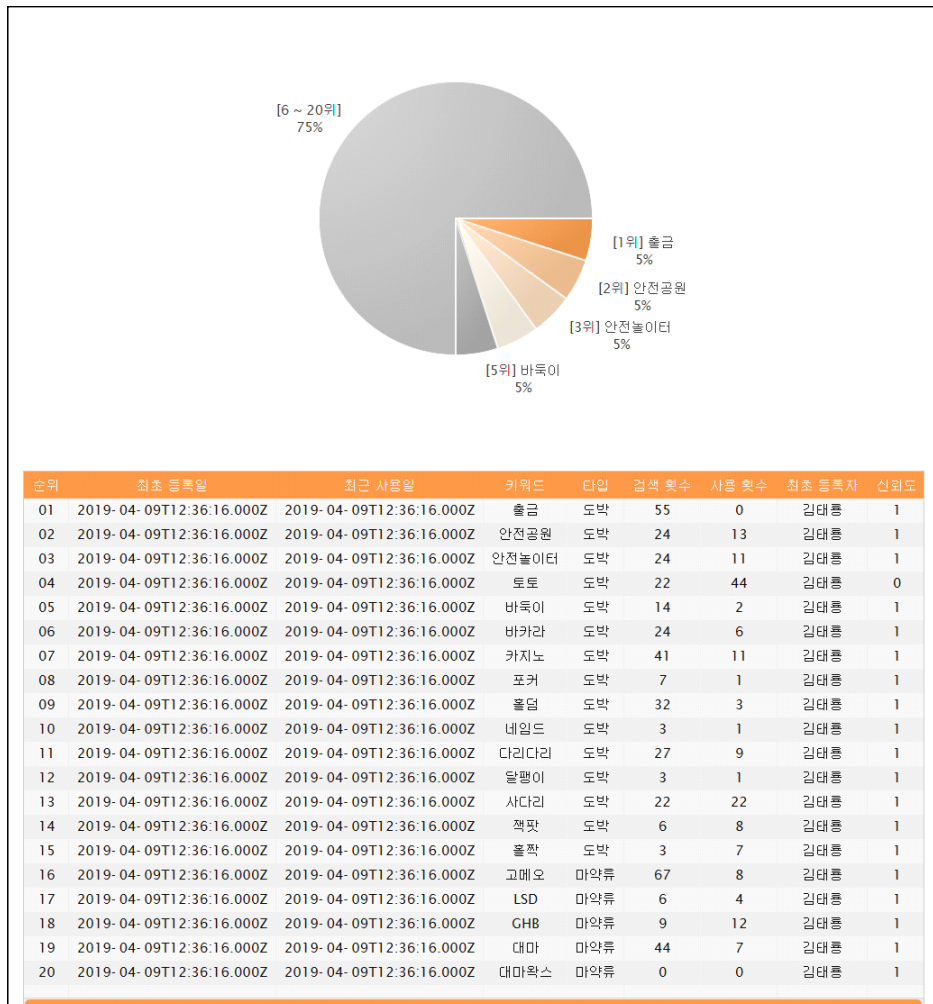
	A	B	C	D	E	F	G	H	I	J
1	최초 탐색	최근 탐색	URL	타입	상위 URL	탐색 횟수	최초 등록자	신뢰도		
2	2019-04-0	2019-04-0	http://060-	음란물	060-900-5	333	김태룡	2		
3	2019-04-0	2019-04-0	http://aekt	음란물	http://aekt	0	김태룡	0		
4	2019-04-0	2019-04-0	http://aktk	연락처	aktkf4x.tur	0	김태룡	0		
5	2019-04-0	2019-04-0	http://aktk	연락처	http://aktk	0	김태룡	0		
6	2019-04-0	2019-04-0	http://aldz	연락처	aldznft.tun	0	김태룡	0		
7	2019-04-0	2019-04-0	http://alsal	음란물	http://alsal	0	김태룡	0		
8	2019-04-0	2019-04-0	http://arae	연락처	araetnp.tur	0	김태룡	0		
9	2019-04-0	2019-04-0	http://aud	연락처	audgpo9.t	0	김태룡	0		
10	2019-04-0	2019-04-0	http://aud	연락처	http://aud	0	김태룡	0		

(NUReport) (URL) (1571841120322)

[그림 4-5] 다운로드 된 CSV 파일

4.2.2 Pie Chart 출력

데이터를 한 눈에 보기 좋게 Pie Chart와 표가 그려진 PNG 이미지 파일로 출력하는 기능을 넣었습니다. 이를 통해 통계 보고서를 작성하여 올릴 경우, Nuricops 회원 및 수사관들에게 요즘 어떤 키워드가 유행하는지, 어떤 키워드가 유해 게시글을 더 잘 잡는지에 대한 정보를 제공하여 유해사이트 모니터링에 도움을 줄 것입니다.



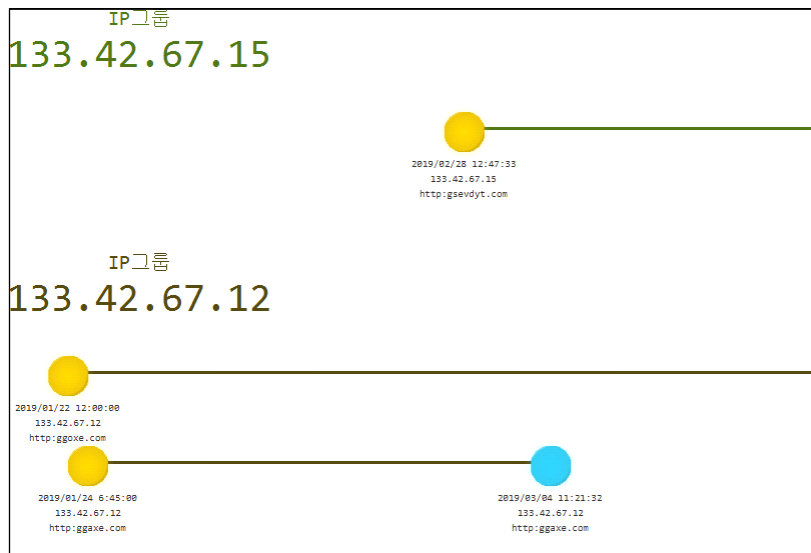
[그림 4-6] 다운로드 된 테스트용 키워드 파이 차트

4.3 시각화

4.3.1 IP주소/URL주소 변경 타임라인

유해사이트의 경우, 수사관의 단속을 피해 IP 및 URL 주소를 이리저리 변경하는 특성을 가지고 있습니다. 때문에 동일 범죄자가 운영 중인 사이트더라도 주소가 다르거나 IP가 다를 경우, 연관성을 찾기 쉽지 않습니다.

이에 변동되는 URL과 IP의 공통점을 찾아 하나의 그룹으로 묶고, 이를 타임라인 형식으로 시각화를 진행하는 기능을 추가하여 IP 주소별로 각 URL의 최초 발견부터 마지막 발견 일자를 쉽게 찾아 볼 수 있도록 하였습니다.

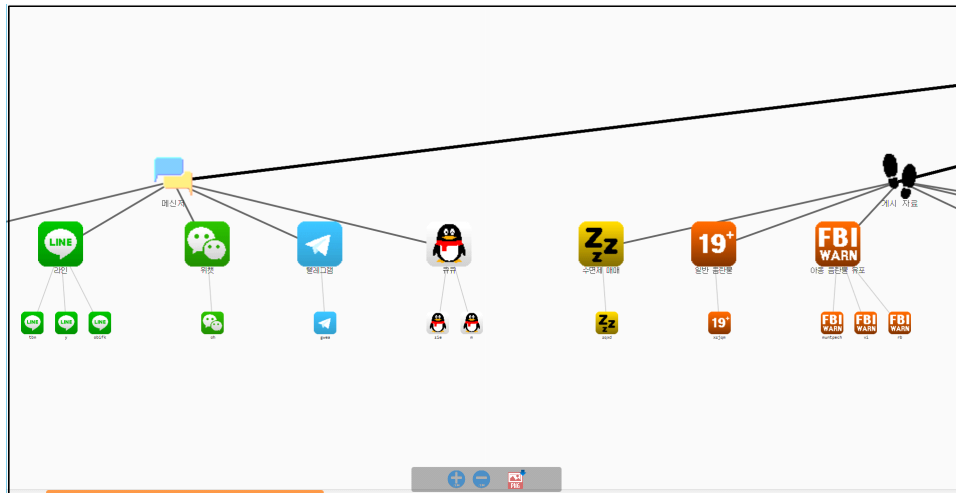


[그림 4-7] 타임라인 형식으로 펼쳐진 주소 변동 내역

4.3.2 유해 사이트 추적 단서 마인드맵

도박 및 마약류의 경우 범인이 판매하는 물품 혹은 운영 중인 사이트, 닉네임, ID 등등이 다양하기 때문에 메모장에 일일이 적어서 읽을 경우, 너무 길어져서 추적 효율이 떨어졌습니다. 때문에 당시 그림판

을 통해 트리 형식으로 나열하였더니, 한 눈에 추적 대상, 단서, 주요 판매 품목들이 눈에 들어왔던 기억을 되살려, 수사관분들에게 도움이 될 수 있도록 수집된 추적 단서를 트리 형식으로 나열하는 기능을 추가하였습니다.



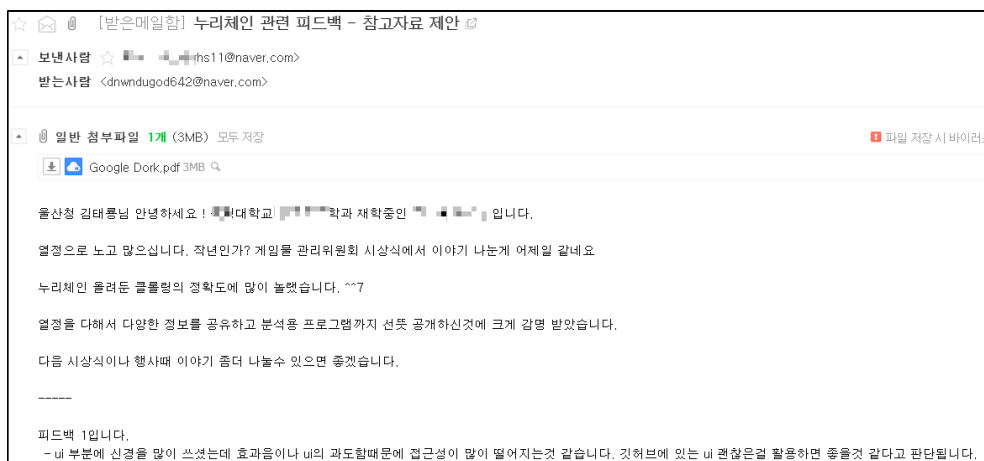
[그림 4-8] 아이콘과 텍스트로 이루어진 트리 시각화

NuriCHAIN alpha 1 시연 영상 : <https://youtu.be/BkHq69RD9I4>

제 5 장 결론

5.1 효과 및 기대효과

아직 수사기관에서 사용한 적이 없기에, 수사대에게 미치는 실질적인 효과를 알 수 없고, 누리캡스 내에서 NuriCHAIN 사용에 대하여 공식적으로 인정하지 않은 상태이기에 모니터링 및 신고율 상승에 대한 수치적인 데이터는 집계 할 수 없는 상태지만, 각 기관에서 NuriCHAIN의 성능을 높게 평가 해 준 덕에 2019년 7월에는 부산지방 경찰청의 한 누리캡스 회원이 간담회에서 NuriCHAIN 시연 설명회를 열었으며, 9월에는 게임물관리위원회로부터 프로그램이 마음에 든다는 연락이 왔습니다. 이외에도 전자 메일을 통해 감사 인사 및 피드백이 자주 오는 모습과 DB에 수집된 데이터를 보면, 적어도 NuriCHAIN은 사람이 직접 하는 것 보다 유해사이트 탐색 시 효율적임을 알 수 있습니다.



[그림 5-1] NuriCHAIN 피드백

또한 현재 NuriCHAIN 크롤러만 배포되었을 뿐, NuReport 데이터 시각화 프로그램은 배포되지 않은 상태이기 때문에, 시각화 프로그램에 대한 실질적 효과도 현재로서는 알 수 없습니다.

이처럼 현재는 실질적인 효과를 알 수 없는 상태이지만, 기대효과는 분명합니다. NuriCHAIN 크롤러가 공식적으로 배포될 경우, 누리캡스 회원들의 모니터링 효율이 상승하여 신고율이 높아 질 것이며, 높아진 신고율 만큼 많은 유해사이트가 차단될 것입니다. 이는 정보통신망 침해범죄 및 정보통신망 이용범죄의 경로가 차단되는 것으로, 전체적인 사이버 범죄율 하락을 기대할 수 있습니다.

또한 ISCR 2019, 2019 사이버안전 학술세미나 등 최신 보안관련 세미나에 참석하여 강연 내용을 들어보면, 늘어나는 사이버 범죄로 인해 이제 국가 기관 혼자서 해결할 수 없고 민간, 학회, 기업이 모두 뭉쳐 해결해야만 한다는 주제가 자주 나옵니다. 이는 NuriCHAIN을 통해 민간 분야에 힘을 보태고, NuReport로 시각화된 자료를 공유하는 것만으로도 전체적인 사이버 치안 향상에 도움이 된다는 뜻이며, 범죄가 일어날 환경을 미리 제거하여 범죄를 방지하는 CPTED 관점으로 바라보아도 “크롤러를 이용해 수사관 및 시민들이 항상 모니터링 중”이라는 환경은 사이버 치안 향상에 크게 도움이 될 것입니다.

5.2 한계 및 향후 연구 과제

NuriCHAIN 및 NuReport는 수사기관에서 이미 사용 중인 크롤러 혹은 시각화 프로그램이 있거나, 필요한 기능이 없다면 본래 제작 목적과는 달리 졸업 작품 정도로만 제출할 수밖에 없는 큰 한계가 있습니다.

또한 실험 결과 및 성능은 만족스러웠지만, 여전히 HTTP를 이용한 통신밖에 할 수 없어서 딥웹(Onion 네트워크) 모니터링을 할 수 없습니다. 경찰 수사를 돕기 위해서는 실제 유해 매체의 근원지인 딥웹 탐색이 필수이기 때문에 2018년도에는 Orchid 를 이용하여 TOR 네트워크에 접속한 뒤, 딥웹을 접속해 보려 하였으나 실패하였습니다. 때문에 딥웹 탐색이 되지 않는 한계는 향후 연구 과제로 남아 있습니다.

이 외에도 현재 누리캡스는 ‘자동신고’를 비허용 하고 있기 때문에, 이용자가 NuriCHAIN 크롤러로 유해사이트를 10000개 찾았다면, 10000

개를 직접 손으로 하나하나 신고해야 합니다. 만일 자동신고를 통한 신고 접수 페이지(혹은 API)가 나오고, 자동신고 점수는 별개로 계산하는 등의 방식을 통해 이 문제가 해결 될 경우, 안드로이드 APP 형식의 NuriCHAIN을 [표 5-1]의 알고리즘대로 제작하여 휴대폰을 사용하지 않는 동안, 백그라운드에서 모니터링을 할 수 있도록 할 것입니다.

순서	내용
1	검색 버튼 클릭.
2	NuriCHAIN 서버로부터 시작 URL을 받아온다.
3	NuriCHAIN 서버로부터 키워드 데이터를 받아온다.
4	시작 URL부터 키워드를 빗대가며 크롤링을 시작한다.
5	얻은 결과물을 서버에 보내고 크롤링을 계속한다.
6	검색 중단버튼 클릭 전 까지 5번 항목을 반복한다.

[표 5-1] NuriCHAIN 앱 구동 알고리즘

‘전 국민 대상의 간편 모니터링 APP’을 목적으로 하고 있기 때문에, 이용자에게 어떠한 유해사이트 URL, 수집 량, 키워드 등의 설정 항목은 나오지 않을 것이며, 앱 구동시 화면 정 중앙에 ‘검색 시작’ 버튼만 보이게 하여 미성년자도 사용할 수 있도록 할 것입니다.

유해 미디어 중, 음란물의 경우 유해성 판단 여부가 복잡합니다. 가슴이 노출되어도 성기가 노출되지 않았더라면 신고할 수 없고, 성기가 보이더라도 의학 지식일 경우 신고할 수 없습니다. 이러한 까다로운 조건들 때문에 신입 누리caps 회원의 경우, 가슴만 노출된 사진을 신고하는 경우가 종종 있었습니다. 때문에 AI를 통한 이미지 분석 기능을 추가하여 유해성 검토 능력을 향상시킬 계획입니다.

유해사이트 신고 시, 다른 누리caps 회원이 신고한 URL은 신고할 수 없는데, 이 때문에 크롤러를 통해 URL을 모두 긁어와도 이 모든 것을 일일이 신고 확인하는 시간이 오래 걸립니다. 따라서 크롤링 이후, 수집된 URL들에 대한 신고여부 확인 기능을 추가시킬 계획입니다.

참 고 문 헌

- [1] 2018 인터넷이용실태조사, 과학기술정보통신부/한국인터넷진흥원, 2019.02.25.
- [2] 이정민 (Lee Jung Min) , 문준섭 (Moon Jun Seob). 2017. 사이버 범죄 예방을 위한 Cyber CPTED 모델 제시에 관한 연구. 한국경찰학회보, 65(0): 23-48
- [3] 누리캡스 홈페이지 - 명예의 전당 / 이달의 누리캡스 -
- [4] 메트로신문 나유리 기자 - 활개치는 '휴대폰 깡'...젊은층 신용불량 나락으로 - (2018-05-28)