https://doi.org/10.33778/kcsa.2023.23.3.029

가짜 광고성 피싱 사이트 탐지 모델 및 대응 기술

이 은 빈*, 조 정 은*, 박 원 형**

요 약

최근 검색 엔진에서의 가짜 광고성 피싱 사이트 노출이 급증하면서 검색 품질 악화 및 개인정보 유출로 인한 피해가 커지고 있다. 특히, ChatGPT와 같은 도구들을 통한 광고성 피싱 사이트 생성이 자동화될 가능성이 증가함에 따라 해당 문제의 심각성은 더욱 빠르게 악화되고 있다. 본 논문에서는 가짜 광고성 피싱 사이트의 소스 코드를 정적 분석하여 구조적 공통점을 도출하였고 그 중 외국 도메인, 리다이렉션을 바탕으로 사이트를 단계적으로 필터링하는 탐지 크롤러를 개발하여 최종적으로 가짜 광고성 게시물들이 탐지됨을 확인하였다. 또한, 가짜 광고성 사이트의 리다이렉션 페이지가 3가지의 유형으로 나뉘어 각 상황에 따라 다른 사이트를 반환하는 것을 검증함으로써 새로운 가이드라인의 필요성을 입증한다. 나아가 기존의 탐지 방법으로 탐지가 불가한 가짜 광고성 피싱 사이트를 대상으로 새로운 탐지 가이드라인을 제안한다.

Detection Models and Response Techniques of Fake Advertising Phishing Websites

Eunbeen Lee*, Jeongeun Cho*, Wonhyung Park**

ABSTRACT

With the recent surge in exposure to fake advertising phishing sites in search engines, the damage caused by poor search quality and personal information leakage is increasing. In particular, the seriousness of the problem is worsening faster as the possibility of automating the creation of advertising phishing sites through tools such as ChatGPT increases. In this paper, the source code of fake advertising phishing sites was statically analyzed to derive structural commonalities, and among them, a detection crawler that filters sites step by step based on foreign domains and redirection was developed to confirm that fake advertising posts were finally detected. In addition, we demonstrate the need for new guidelines by verifying that the redirection page of fake advertising sites is divided into three types and returns different sites according to each situation. Furthermore, we propose new detection guidelines for fake advertising phishing sites that cannot be detected by existing detection methods.

Key words : Fake advertising site detection, Crawling, Malicious domains

접수일(2023년 08월 13일), 수정일(2023년 09월 14일), 게재확정일(2023년 09월 26일)

* 성신여자대학교 융합보안공학과 학부생
* 성신여자대학교 융합보안공학과 학부생
** 성신여자대학교 융합보안공학과 교수(교신저자)

1. 서 론

피싱 사이트는 기본적으로 신뢰할 수 있는 기관인 정부 기관, 금융기관 등의 웹사이트를 사칭하여 접속한 사용자의 개인정보 및 민감정보에 해당하는 각종 정보들을 탈취하는 악의적인 웹사이트를 의미한다[1]. 하지만 기술이 발전함에 따라 피싱 사이트라 정의되는 범주 또한 매우 넓어지고 있다.

최근 인공지능 대화형 챗봇인 'ChatGPT'의 급격한 발전으로 인해 많은 이에게 폭발적인 관심을 얻고 있다. 특히, ChatGPT를 이용해 피싱 사이트 및 가짜 광고성 사이트 제작에 악용되는 사례가 증가하고 있다[2]. ChatGPT를 활용해 제작된 가짜 광고성 사이트의 경우, 특정 국가 및 집단의 트래픽을 유발하는 키워드를 자동으로 추출하여 이를 바탕으로 악의적인 의도의 사이트를 양산할 수 있다[3]. 이러한 고도화된 피싱 사이트로 인해 정보 검색의 부정확성, 개인정보 탈취 등의 피해가 발생하고 있어 이용자들의 불편함이 나날이 증가하고 있다[4, 5].

또한, 기존의 사이트 차단 프로그램 및 서비스가 존재함에도 불구하고 계속해서 새롭게 생성되는 다양한 유형의 피싱 사이트를 막기에는 역부족이라 판단된다. 피싱 사이트의 유형이 다양해지는 만큼 효과적인 탐지를 위해 각 유형에 적합한 맞춤 탐지 및 대응 기법의 필요성이 대두되었다.

따라서 본 논문에서는 기관을 사칭하는 사이트가 아닌, 정상적인 사이트와 같이 키워드 및 문자열이 나열되어 있는 것처럼 보이지만 접속 시에는 유해 광고 및 게임 광고가 나타나는 구조의 현재 새롭게 유행하는 가짜 광고성 사이트를 대상으로 하였다. 이에 대한 대응책을 마련하고자 해당 사이트에 직접 접속하여 구조적으로 공통된 특징을 분석하고 이를 바탕으로 고도화된 피싱 사이트를 탐지 및 대응할 수 있는 기술을 제안하고자 한다.

2. 관련 연구

사준호 외 1명은 피싱 사이트가 사칭하고자 하는 기관의 공식 사이트에 사용자가 접속할 경우, HTTP Referer 헤더 필드 사용으로 피싱 사이트의 URL이 정상 사이트로 유입되는 특성을 이용하여 실시간으로 피싱 사이트를 탐지하는 시스템을 제안하였다. 해당 연구에서는 국내 특정 웹사이트에 탐지 시스템을 적용하여 6일간 40개의 피싱 사이트를 탐지함으로써 HTTP Referer 탐지기법의 효과성을 증명하였다[1]. 백지현 외 1명은 검색 엔진 결과 리스트의 순위를 활용하여 웹 스팸 중 URL 리다이렉션의 탐지 기법을 제안하였다. HTTP 상태 코드, 메타 태그, 스크립트에 의한 리다이렉션을 소개하고 구글의 페이지랭크 알고리즘을 사용하여 URL 리다이렉션 스팸 탐지 알고리즘을 제안하였다[6]. 정연기는 유사 도메인의 규칙성을 기반으로 피싱 범죄용 인터넷 도메인을 반복해서 생성하는 등록자의 특성을 분석하여 50일 동안 활성화된 119개의 피싱 사이트를 판별하였다. 또한, 도메인 등록자 정보를 이용하여 개선된 피싱 사이트 차단 단계별 대응을 제안하였다[7].

3. 가짜 광고성 사이트 분석

본 연구에서는 광범위한 가짜 광고성 사이트 중에서 최근에 양산되는 특정 형태의 광고성 사이트를 탐지 대상으로 한정하였다. 특히, 한국어 웹의 콘텐츠를 무분별하게 크롤링하여 게시글을 생성한 뒤, 인터넷 사용자들이 게시물을 클릭하도록 유도하는 경향을 보이는 가짜 광고성 사이트를 대상으로 연구를 진행하였다. 새롭게 유행하는 가짜 광고성 사이트의 공통된 특성을 도출하기 위한 목적으로 해당 웹사이트들에 직접 접속 후 Burp Suite의 Proxy 기능을 사용하여 특징을 분석하고자 하였다. 해당 기능 사용 시 HTTP GET, POST 요청과 응답을 확인할 수 있어 분석에 더욱 용이하다[8]. 이를 활용해 서버와 클라이언트 간에 주고받는 패킷을 가로채는 스니핑 기법을 사용하였고 HTTP 요청/응답 구조와 속성을 확인할 수 있었다. 앞선 분석을 통해 ChatGPT와 같은 도구로 자동화되어 생성된 웹사이트들은 몇 가지 공통된 특징을 가지고 있음을 확인하였다. 해당 가짜 광고성 사이트들이 공통적으로 보이는 특징은 다음

과 같다.

1) 외국 도메인 사용

한국에서 사용되는 일반적인 도메인 외에 외국 도메인을 사용하는 경향이 있었다. 그 중 .pl(폴란드), .ru(러시아), .fi(핀란드)와 같은 국가 도메인의 사용률이 높음을 확인하였다.

국가 도메인이 아닌 경우, .wiki, .edu 도메인을 사용하거나 드물게는 .com 도메인을 이용하는 경우도 존재했다. 특히 .edu.pl 형태의 도메인을 사용하는 최근 한국어 웹 게시물의 경우에는 높은 확률로 가짜 광고성 사이트인 것을 확인하였다.

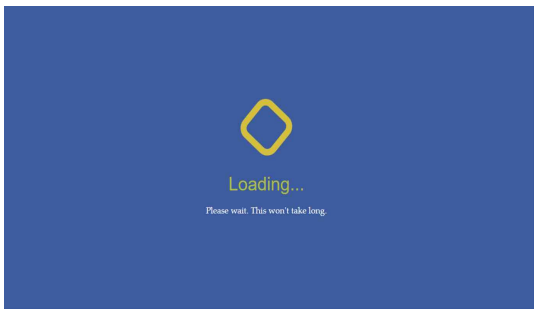
해당 가정을 검증하기 위해 구글 검색 엔진에서 'site:*.edu.pl/' 키워드 검색 및 한국어 웹/최근 1일 게시물 필터링을 적용하였다. 그 결과, 검색 결과로 대부분 가짜 광고성 사이트 게시물이 노출됨을 확인하였다.

2) 연속된 리다이렉션

가짜 광고성 사이트의 URL에 접속하면 다른 URL로 여러 차례의 연속된 리다이렉션이 진행된다. 리다이렉션 동작 방식을 파악하기 위해 본 연구에서는 크롬 개발자 도구(Chrome DevTools)와 Burp Suite의 Proxy 기능을 사용하였다.

검색 엔진에 노출되는 가짜 광고성 사이트 게시물 URL에 접속하면 로딩이 되는 듯한 페이지를 거치게 되는데, 해당 페이지에 삽입되어 있는 스크립트를 통해 또 다른 사이트로의 리다이렉션이 발생한다.

3) 동일한 구조의 로딩 페이지



(그림 1) 동일한 로딩 페이지

공통적으로, 가짜 광고성 사이트에 접속하면 (그림 1)과 같이 동일한 구조를 가진 로딩 페이지로 이동하는 것을 확인하였다. 확인한 가짜 광고성 사이트의 경우, 특정 템플릿 소스가 주로 사용되며 같은 디자인 및 동일한 html 코드를 사용하는 것을 확인하였다.

4) iframe 태그 사용

HTML의 iframe 태그는 외부의 HTML 문서를 현재의 문서에 포함시키는 태그이다. 통상 가짜 광고성 사이트에서 웹 사용자 몰래 리다이렉션 동작을 수행하기 위해 사용되는 경우가 많았다 [9]. 각각의 가짜 광고성 사이트를 통해 접근할 수 있는 로딩 페이지의 소스 코드를 확인해 보면 iframe 태그의 src 속성에 동일한 경로가 삽입되어 있는 것을 확인할 수 있다. 해당 iframe 태그는 아래와 같다.

```
<iframe style="width:5; height:5; display:block; visibility:hidden" id="frmin" src="/media/mainstream/frame.html">
</iframe>
```

(그림 2) 정형화된 iframe 태그

5) 난독화된 script 사용

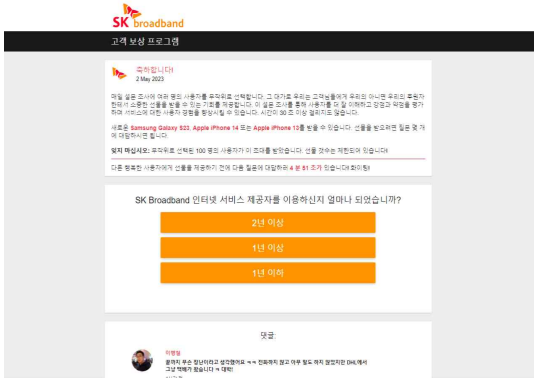
```
function parseURL(_0x2fie49){
  var _0x184cb7=_0x26ed5a;
  try{
    var _0x5eee37=document[_0x184cb7(0x320,'cvVA')]('a');
    return _0x5eee37[_0x184cb7(0x2if,'M@ip')]=_0x2fie49,{
      'source':_0x2fie49,'host':_0x5eee37[_0x184cb7(0x26a,')jmt')],'url':_0x2fie49[
      _0x184cb7(0x159,'N2tT')]+_0x2fie49[_0x184cb7(0x4a2,'j#tF')],'params':(function(){
        var _0x2718c4=_0x184cb7,_0x21ebca={
        },_0x29a5f9=_0x5eee37[_0x2718c4(0x233,'cvVA')][_0x2718c4(0x3f7,'O6b^')](/^\?/,
        '')[_0x2718c4(0x450,'M@ip')]('&'),_0x2bc368=_0x29a5f9[_0x2718c4(0x40d,'9&lO')],
        _0x321a5e=0x0,_0x794b3;
        for(;
        _0x321a5e<_0x2bc368;
        _0x321a5e++){
          if(!_0x29a5f9[_0x321a5e])continue;
          _0x794b3=_0x29a5f9[_0x321a5e][_0x2718c4(0x12d,'2oQE')]('='),_0x21ebca[
          _0x794b3[0x0]]=decodeURIComponent(_0x794b3[0x1][_0x2718c4(0x326,'1lU(')](
          /\+/g,'%20'));
        }
        return _0x21ebca;
      }
      ())
    };
```

(그림 3) HTTP 응답 - 코드 난독화

(그림 3)은 가짜 광고성 사이트의 로딩 페이지 접속 시 반환되는 HTTP 응답 및 HTML 소스 코드 중 스크립트 코드 난독화 부분을 일부 발췌한 것이다. 접속을 시도한 일부 가짜 광고성 사이트에서는 해당 사이트 동작 구조의 원활한 분석을 방해하기 위해 핵심 기능 코드를 난독화하여 나타내고 있다.

실험을 통해 접속한 가짜 광고성 사이트의 경

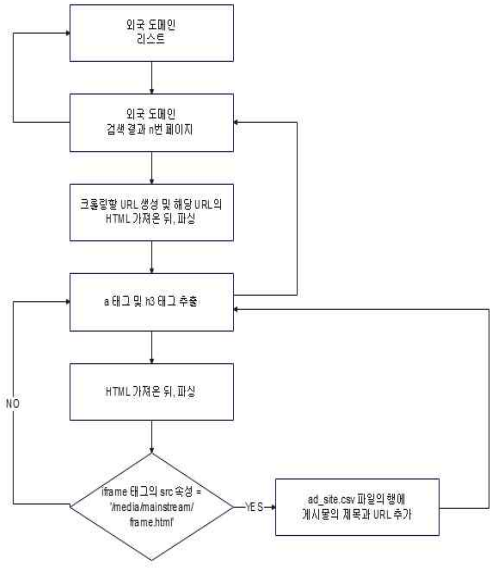
우, 최소 3차례 이상의 리다이렉션을 통해 광고 페이지에 최종적으로 도달했으며, (그림 4)와 같은 광고 페이지도 마찬가지로 정형화되어 있어 동일한 구조의 특정 광고 페이지들이 랜덤하게 노출되는 경우가 많음을 확인하였다.



(그림 4) 최종적으로 리다이렉션된 페이지

4. 피싱 사이트 탐지 크롤러 구현

가짜 광고성 피싱 사이트들의 공통점 및 특징에 대한 검증을 진행하기 위해 피싱 사이트 탐지 크롤러를 개발하였다. 해당 탐지 프로그램의 동작 흐름도와 소스 코드 일부는 다음과 같다.



(그림 5) 탐지 크롤러 흐름도

```
...
# 외국 도메인 리스트
sites = ["edu.pl", "pl"]

# 크롤링할 검색 결과 페이지의 기본 URL
base_url = "https://www.google.com/search?q=site:*.{}/&lr=lang_ko&hl=ko&tbs=lr:lang_1ko,qdr:d&ei=TVB7ZI_EFpXt-AaO1amgBQ&start={}&sa=N&ved=2ahUKEwjPnsLwqaf_AhWVNt4KHY5qClQQ8tMDegQIBhAE&biw=960&bih=964&dpr=1"

# User-Agent 헤더 설정 (코드 비공개)
...
                    iframe_tag = soup2.find('iframe', {'src': '/media/mainstream/frame.html'})  # iframe 태그 검색

# iframe 태그 사용 여부를 확인 후 결과 출력
                    if iframe_tag:
                        rows.append([title.text, link["href"]])  # 행을 리스트로 추가
                writer.writerows(rows)

get_ad_sites(base_url, sites)
```

(그림 6) 탐지 크롤러 소스 코드 일부

본 논문은 가짜 광고성 피싱 사이트 탐지율을 높이기 위하여 기존 연구들의 방법론인 유사 도메인 검색 기법과 웹 소스 코드 분석 기법을 이용하였다. 이전 단계에서 도출한 특징을 바탕으로 해당 특징들을 가진 사이트를 필터링하여 가짜 광고성 사이트의 URL과 검색 결과 페이지 표면에 나타나는 가짜 광고성 게시물의 제목을 수집하는 크롤러를 개발하였다.

4.1 탐지 크롤링 결과

위의 탐지 크롤러를 사용해 가짜 광고성 사이트의 특징들을 상세히 파악하고자 크롤링을 진행

하였다.

1) 외국 도메인 탐지

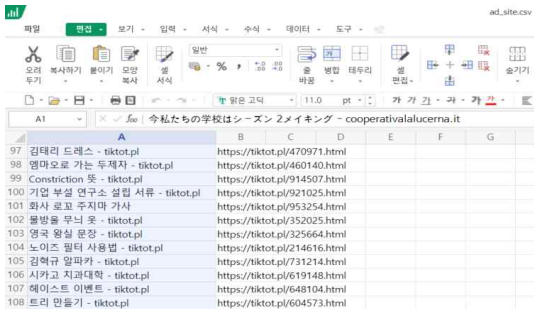
먼저, 외국 도메인을 사용하는 특징으로 일차적인 필터링을 수행한다. sites라는 리스트에 탐지하고 싶은 외국 도메인을 추가하여 도메인별 탐지 반복문을 수행한다. 본 코드에서는 가짜 광고성 피싱 사이트가 자주 탐지되는 도메인인 .edu.pl과 .pl을 선정하여 추가하였다. 구글 및 가짜 광고성 사이트는 크롤링 방지 정책을 적용하고 있으며 그 중 가짜 광고성 사이트는 조금 더 고도화된 정책을 적용하고 있어, 이를 우회하는 User-Agent 헤더 추가 코드가 필요하다. 본 논문에서는 비공개 처리하였다. 최근 게시물의 검색 결과 페이지를 1페이지부터 7페이지까지 순회하기 때문에 크롤링 차단 방지를 위한 딜레이 설정을 추가하였다. 다음으로는 ad_site.csv 파일을 생성하고 csv.writer를 이용하여 csv 파일에 탐지 결과를 출력하도록 한다. base_url에 외국 도메인을 차례대로 대입하여 각각의 도메인 검색 결과 페이지별로 1~7페이지씩 순회한다. 순회하면서 각 게시물을 request.get 메서드를 사용하여 탐지한다.

2) 리다이렉션 탐지

각각의 게시물 URL에 접속하여 만약 iframe의 src 옵션에 /media/mainstream/frame.html이 있으면 가짜 광고성 사이트로 인식하도록 코드를 구성했다. 조건문을 사용하여 가짜 광고성 사이트가 맞으면 csv에 해당 게시물의 제목과 URL을 추가한다. 이전 장에서 추출한 가짜 광고성 사이트의 특징은 포함 관계에 해당하는 특징도 있으므로 피싱 사이트의 가장 확실한 특징으로 여겨지는 iframe src 속성을 탐지의 핵심적인 요소로 정하였다.

3) 코드 실행 결과

한 번의 코드 실행 결과, .edu.pl과 .pl 도메인에서만 100여 개의 가짜 광고성 게시물이 탐지되어 (그림 7)과 같은 결과로 나타났다. 이는 가짜 광고성 피싱 사이트가 만연히 깔려있음을 확인하는 지표라고 볼 수 있으며, 순회 페이지 및 도메인 수를 늘린다면 방대한 양의 가짜 광고성 사이트 수집이 가능할 것으로 예상된다.



(그림 7) 크롤러 코드 실행 결과

5. 피싱 사이트 탐지 및 분석 평가

새롭게 유행하는 가짜 광고성 피싱 사이트는 구글의 탐지를 어렵게 하기 위해 몇 가지 고도화된 방안을 사용하고 있다. 그로 인해 기존의 피싱 사이트를 탐지해오던 방식은 가짜 광고성 사이트를 효과적으로 탐지하기에 불가하다는 한계가 존재한다. 따라서 이러한 기존 탐지 방안의 미비점 보완하여 가짜 광고성 사이트를 탐지할 수 있는 새로운 탐지 모델을 제안한다.

5.1 크롤러 및 구글봇 처리 결과

구글봇(Googlebot)은 구글(Google)의 검색 엔진이 사용하는 웹 크롤러를 칭한다. 구글봇은 웹상의 페이지를 수집하여 구글의 검색 인덱스를 구축하고 업데이트하는 역할을 수행한다. 이를 통해 구글은 사용자가 검색할 때 가장 적합한 검색 결과를 제공할 수 있다[10]. 4장의 가짜 광고성 사이트 탐지 크롤러 개발 연구를 진행하며 리다이렉션의 여부를 피싱 사이트 탐지의 가장 핵심적인 요소로 결정하였다. 그러나 기존에 리다이렉션의 여부를 탐지해 주는 오픈소스의 경우, 가짜 광고성 피싱 사이트의 URL을 리다이렉션을 수행하지 않는 사이트로 판별하는 오류를 범하였다. 이는, 가짜 광고성 사이트가 해당 사이트에 접근하는 주체가 사람인지 아닌지에 대한 필터링이 적용되어 있기 때문에 발생한 문제이다. 가짜 광고성 페이지의 경우, 해당 페이지에 접속하는 주체를 구분하는 필터링을 통해 사람과 사람이 아닌 대상을 구분하고 그 결과 인식한 대상에 따라 상이한 페이

지를 반환한다.

5.1.1 정상 사용자로 인식될 시

일반적으로 구글 검색 결과 페이지에서 일반 사용자가 게시물을 클릭하는 경우, 정상 사용자로 인식된다. 정상 사용자로 인식될 경우, (그림 1)과 같은 로딩 페이지가 출력되며 몇 번의 리다이렉션을 통해 최종적으로 가짜 광고성 피싱 사이트로 접속된다. 본 논문에서 개발한 탐지 크롤러는 크롤링 차단 우회 기법을 통해 크롤러가 아닌 정상 사용자로의 접근이 가능하였다.

5.1.2 구글봇으로 인식 시

구글 검색 시 게시물의 미리보기로 노출되는 게시물 제목 및 관련 내용이 존재하는 정상 사이트로 접속된다. 이러한 수법은 해당 사이트를 정상 사이트로 보이게 하여 가짜 광고성 사이트로 분류되는 것을 피하기 위한 작업이다. 이때, 정상 사이트는 광고가 나타나는 페이지가 아닌 게시물의 내용이 나타나는 페이지를 의미한다. 이는 정상 사이트처럼 작동하나 시각적으로 의미 없는 단어들의 나열로 인해 정상적인 게시물로 볼 수 없다.

(그림 8) 구글봇으로 인식 시 노출되는 페이지

5.1.3 봇으로 의심될 시

(그림 9) 봇으로 의심될 시 노출되는 페이지

파이썬의 requests 라이브러리를 사용하거나 User-Agent 조작 우회 기법만을 사용할 경우, 해당 사이트에서는 사용자를 봇으로 의심하고 봇이 아님을 확인하기 위해 버튼을 클릭하도록 설정하여 크롤링을 방지하고 있다.

5.2 기존 연구와의 차별점

기존 URL 리다이렉션을 탐지하는 방법으로 통용되던 몇 가지 특성은 HTTP 상태 코드, 메타 태그, 스크립트 태그 등이 있다. 하지만 최근 유행하고 있는 가짜 광고성 피싱 사이트는 해당 특징에 따른 탐지법이 대다수 적용되지 않는다.

기존의 URL 리다이렉션이 수행되는 사이트의 경우에는 300번 대의 HTTP 상태 코드를 사용하는 경향이 있었다. 따라서 해당 URL 요청 시, 300번 대의 상태 코드를 반환한다면 리다이렉션이 수행되는 것으로 판단하여 탐지가 가능하였다. 하지만 본 논문의 피싱 사이트의 경우 200번 대의 HTTP 상태 코드를 반환함을 확인하였다.

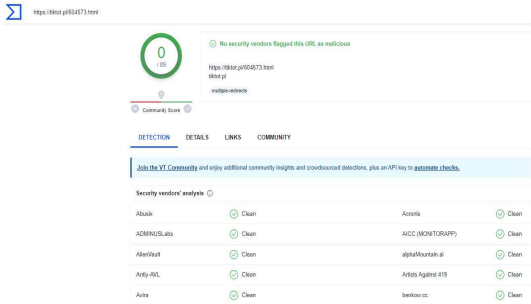
현재 구글에 'URL Redirection Detection'을 검색하면 출력되는 대부분의 여러 탐지 서비스들 역시 HTTP 상태 코드를 이용한 탐지를 수행하는 것을 확인하였으며, 본 연구의 대상이 되는 가짜 광고성 피싱 사이트를 탐지해본 결과 역시 URL 리다이렉션이 일어나지 않는 사이트로 분류하는 것을 확인하였다.

또한 메타 태그나 스크립트 태그 사용 여부를 통한 탐지 알고리즘을 작성하여 실행한 결과, 유의미한 결과를 도출해 내지 못함을 확인하였다.

5.3 제안하는 피싱 사이트 차단 기술

앞 장에서 제안한 탐지 기술을 바탕으로 크롬의 확장 프로그램 형식의 차단 기술을 구현하면 사용자의 편의성 차원에서 큰 시너지를 낼 것으로 예상한다. 예를 들면, 크롬 확장 프로그램 중 uBlacklist는 사용자가 수동으로 차단할 도메인을 설정하거나 각각의 게시물 옆에 위치한 [차단하기] 버튼을 이용하여 사이트 차단을 수행한다[11]. 이와 같은 기능을 사용하면 사용자의 자율성 보장과 가짜 광고성 피싱 사이트 자동 차단이 가능할 것이다. 또한, 크롬 브라우저를 사용하였을 시 굳이 프로그램을 켜지 않아도 즉각적으로 차단할 수 있다는 점에서 뛰어난 접근성을 보인다. 이러한 자동화된 차단 기법을 통해 일차적인 필터링을 수행한

후에 사용자가 직접 발견해 낸 가짜 광고성 피싱 사이트가 있을 시, 구글에 신고할 수 있는 리포트 기능을 추가한다면 검색 결과 오염 및 품질 악화를 개선하는 데 긍정적 영향을 끼칠 수 있을 것이다.



(그림 10) 제안하는 피싱 사이트 바이러스 검사

위 그림은 해당 광고성 피싱 사이트를 Virus Total에 URL 검색한 결과이다. 해당 사이트가 악성코드를 내포하고 있을 확률은 높지 않으나 검색 품질 악화를 초래하며 많은 이들이 불편을 겪고 있는 점을 감안했을 때 하루빨리 해결되어야 할 문제임이 자명하다.

6. 결 론

최근 전 세계를 통틀어 최고의 검색 엔진이라 여겨지는 구글의 키워드 검색 결과의 신뢰성이 점차 하락하고 있다. 이러한 검색의 품질 악화는 가짜 광고성 사이트가 검색 엔진을 통해 전혀 걸러지지 않은 채 결과로 노출되면서 야기되었다[12].

본 연구는 가짜 광고성 사이트가 지속적으로 생성되고 검색 결과로써 노출되는 현상을 개선하고자 하는 목적으로 수행되었다. 심각한 검색 결과 오염을 유발하는 가짜 광고성 사이트 간의 유사도를 파악하여 이에 대한 탐지 및 대응 방안을 제안하고자 가짜 광고성 사이트에 직접 접속하여 해당 사이트의 소스 코드를 정적 분석함으로써 유사도가 높은 사이트의 공통된 포맷을 도출하였다.

나아가 분석한 공통적 특징을 바탕으로 가짜 광고성 사이트의 특징들을 단계적으로 필터링하여 탐지하는 크롤러를 개발함으로써 검색 결과 페이지에 노출된 가짜 광고성 사이트의 URL과 게시글 제목을 수집하였다. 실험에서 특정 도메인을 선정하여 코드를 실행해본 결과, 한 번의 실행 시 100여 개의 가짜 광고성 사이트가 성공적으로 탐지되었음을 검증하였다. 해당 탐지 크롤링의 결과로, 가짜 광고성 사이트의 게시물 제목과 URL을 csv 파일의 데이터셋으로 생성함으로써 추후 진행되는 가짜 광고성 사이트 탐지 및 차단 연구에 기여할 수 있으리라 생각한다. 또한 3, 4장에서의 탐지 크롤러 연구를 수행하면서 이와 유사한 구조의 다른 가짜 광고성 사이트도 존재함을 확인하였다. 본 연구에서 진행한 도메인, 리다이렉션 필터를 다른 가짜 광고성 사이트의 특성에 맞게 적용한다면 이 역시 탐지 가능성이 있을 것이라 기대된다. 본 연구를 통해 검색 결과의 오염이 완화되고 더욱 효율적인 탐지 및 대응 기술이 발전하게 된다면 궁극적으로 가짜 광고성 피싱 사이트로 인한 금전적 손실 및 개인정보 유출과 같은 직·간접적인 피해 예방에 기여할 수 있을 것이다.

참고문헌

[1] 사준호, 이상진, "피싱사이트 실시간 탐지 기법", Journal of the Korea Institute of Information Security & Cryptology, 2012.08.31.

[2] 김혜경, "피싱 사이트 만들어줘"…해커들에게 악용되는 챗GPT, 아이뉴스24, https://www.inews24.com/view/1565908, (검색일: 2023.06.07.).

[3] 전수한, '음지'부터 확산되는 챗 GPT, 피싱사이트·사이버범죄 악용, 문화일보, https://www.munhwa.com/news/view.html?no=2023022101070121313001, (검색일: 2023.06.07.).

[4] 전수한, "OOO처럼" 명령하면 똑같은 '피싱사이트'…공범되는 AI, 문화일보, https://www.munhwa.com/news/view.html?no=2023022101070921313002, (검색일: 2023.06.07.).

[5] 김대은, '에스파 카리나' 검색했는데 결과가...알고도 당하는 낚시, 범인은 이놈? [아이티라떼], 매일경제, https://www.mk.co.kr/news/it/10742721, (검색일: 2023.06.07.).

[6] 백지현, 김성권, “URL 리다이렉션 스팸 탐지 기법”, Proceedings of the Korean Information Science Society Conference, 2007.10.26.
[7] 정연기, “피싱사이트들의 도메인 특성을 이용한 사이버 범죄 차단 방안”, 디지털포렌식연구, 2015.
[8] Googlebot, https://developers.google.com/search/docs/crawling-indexing/googlebot?hl=ko.
[9] 장진호, 방진숙, “웹서비스 리소스 캐싱에 따른 iframe의 성능 비교 연구”, 지식과 교양, 2021.
[10] Modifying HTTP requests with Burp Proxy, https://portswigger.net/burp/documentation/desktop/getting-started/modifying--http-requests.
[11] uBlacklist, https://chrome.google.com/webstore/detail/ublacklist/pncfbmialoiaghdehhbnbhkkgmjanfhe.
[12] 정용인, “구글 한글검색 품질, 왜 나빠졌나”, 주간경향, https://weekly.khan.co.kr/khnm.html?mode=view&artid=202303241250581&code=114, (검색일: 2023.06.07.).

〔저자소개〕

이 은 빈 (Eunbeen Lee)

2020년 3월 ~ 현재 성신여자대학교 학사과정
email : eun3in1ee@gmail.com

조 정 은 (JeongEun Cho)

2020년 3월 ~ 현재 성신여자대학교 학사과정
email : 20200958@sungshin.ac.kr

박 원 형 (WonHyung Park)

2002년 서울과학기술대학교 산업정보시스템공학과 공학사
2006년 서울과학기술대학교 정보산업공학과 공학석사
2009년 경기대학교 정보보호학과 이학박사
2015년 성균관대학교 컴퓨터교육학과 박사수료
2020년 ~ 2022년 상명대학교 정보보안공학과 부교수
2023년 3월 ~ 현재 성신여자대학교 융합보안공학과 부교수
email : whpark@sungshin.ac.kr