

■ (네이버 후원) 2016년도 학생논문 경진대회 수상작

HTML 태그 순서를 이용한 불법 사이트 탐지 자동화 기술

(An Automated Technique for Illegal Site Detection
using the Sequence of HTML Tags)

이 기룡 [†]

(Kiryong Lee)

이희조 ^{**}

(Heejo Lee)

요약 2001년 비트토렌트 프로토콜이 설계된 후로 음악, 영화, 소프트웨어 등 모든 것을 다운로드할 수 있게 되었다. 이를 통해 저작권이 있는 파일이 무분별하게 공유가 되었고 저작권자들은 많은 피해를 입었다. 이 문제를 해결하기 위해 국가에서는 관련법을 제정하였고 ISP는 불법 사이트를 차단하였다. 이러한 노력들에도 불구하고 pirate bay와 같은 불법 사이트들은 도메인을 바꾸는 등 쉽게 사이트를 재오픈하고 있다. 이에 우리는 재오픈된 불법 사이트를 쉽게 탐지하는 기술을 제안한다. 이 자동화 기술은 구글 검색엔진을 이용하여 도메인을 수집하고, 최장공통부분수열(LCS) 알고리즘을 이용하여 기존 웹페이지 태그와 검색된 웹페이지 태그를 비교, 유사도를 측정한다. 실험을 위해 총 2,383개의 검색 결과를 구글 검색으로 얻었다. LCS 유사도 알고리즘을 적용하여 검사한 결과 44개의 해적 사이트를 탐지하였다. 또한 해외 불법 사이트에 적용한 결과 805개 검색 도메인에서 23개의 불법 사이트를 탐지하였다. 이를 통해 제안된 탐지 자동화 기술을 사용한다면 불법 사이트가 재 오픈을 하더라도 쉽게 탐지할 것으로 보인다.

키워드: 불법 사이트, 최장 공통 부분 수열, 자동화 탐지 기술, HTML 태그

Abstract Since the introduction of BitTorrent protocol in 2001, everything can be downloaded through file sharing, including music, movies and software. As a result, the copyright holder suffers from illegal sharing of copyright content. In order to solve this problem, countries have enacted illegal share related law; and internet service providers block pirate sites. However, illegal sites such as pirate bay easily reopen the site by changing the domain name. Thus, we propose a technique to easily detect pirate sites that are reopened. This automated technique collects the domain names using the google search engine, and measures similarity using Longest Common Subsequence (LCS) algorithm by comparing the tag structure of the source web page and reopened web page. For evaluation, we collected 2,383 domains from google search. Experimental results indicated detection of a total of 44 pirate sites for collected domains when applying LCS algorithm. In addition, this technique detected 23 pirate sites for 805 domains when applied to foreign pirate sites. This experiment facilitated easy detection of the reopened pirate sites using an automated detection system.

Keywords: illegal site, longest common subsequence algorithm, pirate site, automated detection technique, HTML tag

[†] 학생회원 : 고려대학교 컴퓨터.전파통신공학과
krlee@korea.ac.kr

심사완료 : 2016년 6월 29일
(Accepted 29 June 2016)

^{**} 종신회원 : 고려대학교 컴퓨터.전파통신공학과 교수(Korea Univ.)
heejo@korea.ac.kr
(Corresponding author)

Copyright©2016 한국정보과학회 : 개인 목적이나 교육 목적인 경우, 이 저작물의 전체 또는 일부에 대한 복사본 혹은 디지털 사본의 제작을 허가합니다. 이 때, 사본은 상업적 수단으로 사용할 수 없으며 첫 페이지에 본 문구와 출처를 반드시 명시해야 합니다. 이 외의 목적으로 복제, 배포, 출판, 전송 등 모든 유형의 사용행위를 하는 경우에 대하여는 사전에 허가를 얻고 비용을 지불해야 합니다.
정보과학회논문지 제43권 제10호(2016. 10)

논문접수 : 2016년 4월 20일
(Received 20 April 2016)
논문수정 : 2016년 6월 14일
(Revised 14 June 2016)

1. 서 론

2001년 비트토렌트 프로토콜이 설계된 후로 음악 파일 뿐만이 아니라 영화, 소프트웨어 등 컴퓨터에서 이용 가능한 모든 것을 공유를 통해 다운로드 할 수 있게 되었다. 이를 통해 저작권이 있는 파일이 무분별하게 공유가 되었고 그 결과 저작권자들은 많은 피해를 입었다. 2011년 소프트웨어 판매에서 미국은 전체 판매의 약 19%, 한국의 경우는 약 40%, 중국의 경우 약 77%의 피해를 입는 등 전 세계적으로 불법 파일이 무분별하게 공유되고 있다[1].

이러한 문제를 해결하기 위해 국가에서는 관련법을 제정하거나 Internet Service Provider(ISP)에서 사이트를 차단하는 등 피해를 줄이기 위해 많은 노력을 하였다. 스웨덴의 경우 2009년 IPRED 법을 개정하여 불법 다운로드 트래픽의 16% 감소와 음원 판매를 36% 올렸다[2]. 영국은 pirate bay와 다른 주요 불법 저작권 파일 배포 사이트 19개를 막음으로써 netflix와 같은 합법 스트리밍 사이트의 이용률이 평균 12% 증가했다[3]. 한국의 경우, 방송통신심의위원회에서 2015년에 약 1만 6천여 개의 불법 침해 사이트를 차단하였다[4]. 이러한 노력들에도 불구하고 불법 사이트들은 해외로 도메인을 바꾸는 등 쉽게 사이트를 재 오픈하며 악의적인 행동을 계속하고 있다.

이에 우리는 해적 사이트와 같은 불법 사이트들이 차단된 후 사이트를 재 오픈하더라도 자동으로 탐지하는 기술을 제안한다. 구글 검색엔진을 이용하여 도메인을 수집한 뒤, 두 HTML 태그간의 유사도를 측정하여 기존에 차단되었던 사이트의 웹페이지 태그 순서와 재 오픈한 사이트의 웹페이지 태그 순서를 비교하여 유사도를 측정하였다.

실험을 위하여 한국 상위 50개(Alexa 2015년 08월 23일 기준) 해적 사이트에서 웹페이지 소스 코드를 수집하였고, 약 한달 후인 2015년 9월 30일에 각각의 사이트 이름을 이용하여 구글에서 도메인을 다시 수집하였다. 실험 결과 2,383개의 검색된 도메인 중 총 44개의 해적 사이트를 찾아내었다. LCS 유사도 측정 알고리즘을 적용하였을 때 44개의 해적 사이트 모두 탐지하였고, 임계치가 80%일 때 오탐율과 미탐율이 0이었다. 또한 외국 상위 10개 해적 사이트를 대상으로 실험을 한 결과 805개의 검색 도메인 중 23개의 해적 사이트를 탐지 할 수 있었다.

이를 통해 불법 사이트들이 차단 후 재 오픈하였을 때 기존에 사용한 웹페이지 소스를 그대로 사용한다는 것이 확인되었다. 또한 제안된 탐지 자동화 기술을 이용한다면 이러한 불법 사이트들에 의한 피해를 최소화 할

수 있을 것으로 예상된다.

본 논문의 구성은 다음과 같다. 1장의 서론에 이어 2장에서는 웹 페이지 유사도를 측정에 관한 관련 연구에 대해 살펴보고, 3장에서는 본 논문에서 제안하는 해적 사이트 탐지 방법에 대하여 설명한다. 4장에서는 이를 이용한 실험과 그 결과에 대해 분석하며 마지막으로 5장에서 결론을 도출한다.

2. 관련 연구

불법 사이트들을 자동으로 탐지하는 방법은 아직 없으며, 이러한 사이트를 찾기 위해서는 신고에 의하거나 하나씩 수동으로 찾을 수밖에 없었다. 그러나 제안할 시스템이 유사도를 이용한다는 점에서 피싱 사이트를 탐지하는 방법과 비슷하기 때문에 피싱 사이트 탐지에 관한 연구 몇 가지를 소개한다.

Angelo 등은 레이아웃 구조 유사성을 이용하여 피싱 사이트를 탐지하였다[5]. 그들은 피싱 페이지 탐지 프로그램을 만들어 사용자가 비밀번호와 같은 양식을 입력할 때 “watch list”에 있고 도메인이 일치하면 DOM 트리 구조를 이용하여 유사도를 측정하였다. 그러나 이 연구는 사용자가 직접 웹에 접근을 하고, 양식을 입력하였을 때 검사를 하므로 사전에 차단할 수 없다.

Roopak와 Tony는 피싱 사이트 탐지를 구글 검색과 HTML 코드를 이용하여 측정하였다[6]. 그들은 유저가 접속한 웹 페이지의 타이틀과 5개의 용어 빈출도를 기반으로 구글에 검색한 후 그 결과 url들의 HTML 코드의 코사인 유사도를 측정하여 비교를 하였다. 그러나 해적 사이트의 경우 메인 페이지의 태그는 잘 바뀌지 않고 컨텐츠가 많이 바뀌므로 HTML 코드만을 비교하면 유사도가 낮아질 수 있다. 따라서 불법 사이트 탐색에는 태그를 비교하는 것이 더 높은 유사도를 얻을 수 있다.

Maurer와 Herzner는 웹페이지의 유사도를 화면간의 픽셀을 이용하여 비교하였다[7]. 그들은 피싱 페이지와 원본 페이지의 화면을 저장한 뒤 두 화면의 픽셀을 비교하였다. 그러나 불법 사이트의 경우 수익이 대부분 광고에서 창출되고 매 시간마다 다른 광고가 나타나기 때문에 픽셀의 변화가 많아 픽셀 비교기법의 방식에서는 유사도가 낮을 수 있다. 또한 불법 사이트의 경우 대부분 코드를 재사용하기 때문에 화면 간의 비교를 하지 않아도 높은 유사도를 얻을 수 있다.

3. 불법 사이트 탐지 기술

3.1 기술 개요

그림 1은 탐지 기술의 작동 순서를 보여준다. 시스템의 작동 순서는 5단계로 표현되는데, 먼저 기존에 가지고 있던 차단된 불법 사이트의 이름을 구글에 검색하여

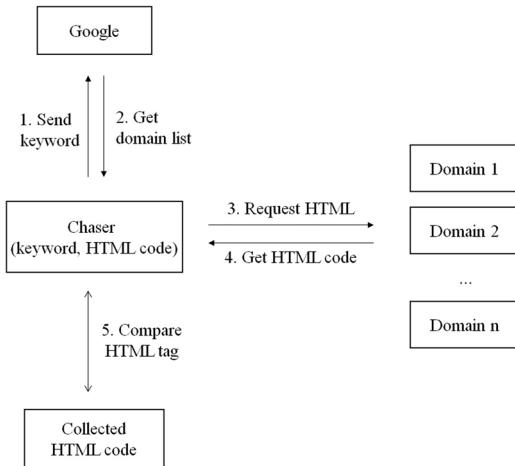


그림 1 기술 개요

Fig. 1 Technique overview

도메인 리스트를 획득한다. 그 다음 획득한 도메인 리스트를 이용하여 각각의 도메인에 접속한 뒤 메인 페이지의 HTML 코드를 획득한다. 마지막으로 기존에 가지고 있던 불법 도메인들의 웹페이지 소스 코드의 태그 집합과 검색으로 새롭게 얻은 웹페이지 소스 코드의 태그 집합을 유사도 측정 알고리즘을 통하여 유사도를 측정한다.

3.2 구글 검색

제안된 기술은 구글 검색을 이용하는데, 구글은 PageRank 알고리즘을 사용한다[8]. PageRank 알고리즘은 단순히 검색 인덱스를 하는 것이 아니라 다른 페이지와의 링크의 수를 이용하여 중요 사이트를 상위에 보여준다. 이러한 특징 덕분에 불법 사이트의 이름만으로 검색을 하더라도 재 오픈한 사이트를 쉽게 검색할 수 있다. 또한 구글이 제공하는 검색 옵션 중 필터 옵션을 사용하면 중복된 도메인에서의 결과를 최소화해 주기 때문에 적은 노력으로도 대부분의 사이트를 검색할 수 있게 된다.

3.3 LCS 알고리즘을 이용한 유사도 측정

웹페이지의 소스 코드를 얻었다면 다음으로 할 일은 기존에 가지고 있던 차단된 사이트의 웹페이지 태그와 검색으로 얻은 웹페이지 태그의 유사도를 비교하는 것이다. 웹페이지 전체를 유사도 측정에 이용하지 않는 이유는 해적 사이트와 같은 불법 사이트에는 구조는 일정하지만 웹페이지 내의 컨텐츠의 내용이 수시로 바뀌기 때문이다. 그렇기 때문에 웹페이지 전체의 유사도를 측정한다면 그만큼 탐지율이 떨어지게 되고, 웹페이지의 크기가 크기 때문에 비교 속도도 더 느려지게 된다.

유사도를 측정하기 위해서는 먼저 기존에 가지고 있

```

Input : Two Sequence of HTML Tag
Output : Result of Similarity Measurement
Function LCS_Similarity(X[1..m], Y[1..n])
{
    C = array[0..m, 0..n]      // create array
    for i = 0; i <= m; i++      // initialize array
        C[i, 0] = 0
    for j = 0; j <= n; j++
        C[0, j] = 0
    for i = 1; i <= m; i++      // LCS algorithm
        for j = 1; j <= n; j++
            if X[i] == Y[j]
                C[i, j] = C[i-1, j-1] + 1
            else
                C[i, j] = max(C[i, j-1], C[i-1, j])
    return min(C[m, n] / m, C[m, n] / n)
}
  
```

그림 2 LCS 유사도 측정 알고리즘

Fig. 2 LCS similarity measurement algorithm

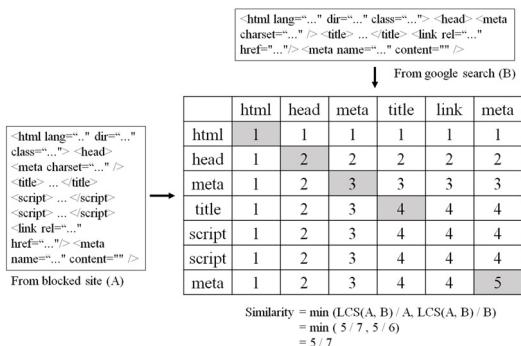


그림 3 LCS 유사도 측정 예

Fig. 3 Example of LCS similarity match

던 불법 도메인의 웹페이지 태그와 검색으로 얻어낸 웹페이지의 태그를 추출한다. 그 다음 추출된 각 페이지의 태그들을 LCS 알고리즘을 적용하여 LCS 유사도 값을 구한다. 마지막으로 얻어진 LCS 값을 각 웹페이지의 태그 개수로 나누어 더 작은 값을 결과 유사도로 출력한다.

그림 2는 유사도 측정 알고리즘을 보여준다. 알고리즘의 입력 값으로는 두 웹페이지의 태그 리스트가 들어가며, 두 웹페이지의 유사도가 출력 값으로 나온다. 예를 들어 두 웹페이지의 태그 개수가 각각 7개(html, head, meta, title, script, script, meta)와 6개(html, head, meta, title, link, meta)이고, 태그 간의 LCS 값이 5라면 이들의 LCS 유사도는 5/7과 5/6의 최솟값인 5/7이 된다. 이 과정을 그림 3에 표현하였다.

4. 실험 결과

4.1 실험데이터

실험을 위하여 우리는 Alexa 2015년 8월 23일 한국 기준 상위 50개 해적 사이트를 탐지 대상으로 정하여

표 1 Alexa 2015년 8월 23일 기준 상위 50개 해적 사이트

Table 1 Top 50 pirate sites on Alexa Aug. 23, 2015

번호	이름 (중복개수)	국내 순위	번호	이름 (중복개수)	국내 순위
1	보고**	29	16	토렌트**	704
2	이토**	38	17	av** (2)	873
3	토렌트** (2)	51	18	av**	1924
4	토** (4)	54	19	빼큐토** (2)	2008
5	티** (2)	59	20	졸라**	3008
6	토렌트** (3)	97	21	T**	3380
7	티** (2)	146	22	토** (3)	3454
8	토렌트** (2)	209	23	find**	6117
9	마루**	230	24	티** (3)	6579
10	토렌**	276	25	토렌트**	10152
11	토렌**	350	26	xx** (3)	N/A
12	오토** (2)	389	27	에스**	N/A
13	야**	400	28	다운로**	N/A
14	토렌트** (3)	407	29	토렌**	N/A
15	티**	626	30	빅**	N/A

각각의 사이트에 대한 웹페이지 소스를 추출하였다. 상위 50개 해적 사이트 중 일부는 같은 이름을 사용하고 있고, 중복을 제외하면 30개의 해적 사이트가 있었다. 표 1은 50개 해적 사이트의 이름과 Alexa 기준 국내 순위에 대한 요약이다.

이들이 재 오픈한다는 것을 확인하기 위해 약 2주 후인 2015년 9월 6일에 50개의 사이트들을 재 추적한 결과 차단되지 않은 사이트가 27개, 차단된 사이트가 14개, 재 오픈한 사이트가 14개 있었다. 이들 중 다운로**은 서비스를 중단하였고, 나머지 29개의 사이트는 차단이 되더라도 재 오픈을 계속 하는 것이 확인되었다. 또한 재 오픈을 할 때 기준에 사용하였던 웹페이지 소스 코드를 그대로 사용함을 알 수 있었다. 이를 토대로 2015년 9월 30일에 제안된 방법을 토대로 실험을 진행하였다.

4.2 구글 검색 결과

탐지에 있어 가장 중요한 요인은 비교 대상이 되는 도메인을 찾는 것이다. 그러나 검색된 모든 도메인을 비교하면 큰 시간이 걸리므로 한정된 양의 검색 결과로 실험을 진행하는 것이 좋다. 이에 우리는 실험에 사용할 30개의 해적 사이트 이름으로 검색하여 검색 결과의 평균 도메인 개수를 측정하였다.

그림 4는 구글에 검색을 하였을 때 얻어지는 도메인의 개수를 나타낸다. 필터를 적용하지 않았을 때에는 100개의 검색 결과 중 중복을 제외한 도메인 개수가 평균 26개였다. 그러나 필터를 적용하였을 때에는 중복 도메인이 대부분 검색되지 않기 때문에 도메인 개수의 평균이 75개였다. 또한 300개 검색 결과로 봤을 때에는 필터를 적용하지 않았을 때 62개, 적용하였을 때 108개로 일정 수준 이상이 되면 도메인 개수의 최대치가 크게 증가하

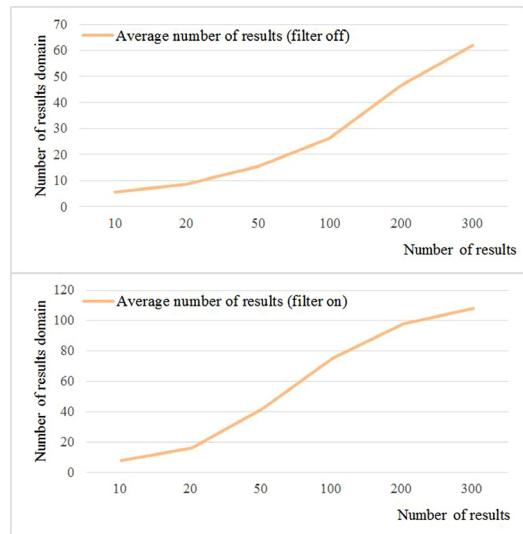


그림 4 필터에 따른 검색 결과 도메인의 개수
Fig. 4 Number of result domains applying filter on/off

지 않음을 알 수 있었다. 따라서 필터 옵션을 적용한다면 검색 결과를 100개로 하여도 대부분의 도메인을 얻을 수 있었기에 필터를 적용한 뒤 100개의 결과만으로 나머지 실험을 진행하였다.

4.3 국내 상위 50개 해적 사이트 실험 결과

2015년 9월 30일에 29개의 해적 사이트 이름으로 구글에 검색한 결과 2,383개의 도메인을 추출하였다. 이들 2,383개의 도메인에 대하여 각각 수동으로 검사해 본 결과 총 44개의 재 오픈 / 미러 사이트가 있었다. 이 결과를 바탕으로 코사인 유사도와 LCS 유사도 측정 알고리즘을 이용하여 얼마나 많은 사이트를 탐지하는지 확인하였고 이에 대한 결과를 표 2와 표 3에 요약하였다.

표 2 코사인 유사도 측정 결과(국내)

Table 2 Result of cosine similarity measurement (domestic)

	Threshold (%)					
	50	60	70	80	90	95
True Positive	44	44	44	44	44	44
False Positive	1400	1171	862	508	140	26
False Negative	0	0	0	0	0	0

표 3 LCS 유사도 측정 결과(국내)

Table 3 Result of LCS similarity measurement (domestic)

	Threshold (%)					
	50	60	70	80	85	90
True Positive	44	44	44	44	44	43
False Positive	37	5	4	0	0	0
False Negative	0	0	0	0	0	1

코사인 유사도의 경우 유사도 임계치의 값이 높을수록 오탐이 줄어드는 것을 볼 수 있다. 그러나 95%의 임계치를 기준으로 측정하였을 때에도 26개의 오탐을 보였다. 이는 코사인 유사도의 경우 태그의 출현 빈도의 벡터 값만을 결과 값으로 내는데 HTML의 경우 태그의 종류가 한정되어 있고 동일 태그의 수만 비슷하다면 높은 유사도를 내기 때문인 것으로 나타났다.

반면 LCS 유사도의 경우 유사도 임계치를 50%로 하였을 때에는 오탐이 37개로 약 1%의 오탐율을 보였다. 또한 유사도 임계치 80% 이상일 때 오탐이 없었다. 하지만 임계치 90% 이상일 때 LCS의 경우 1개의 미탐이 존재하는데, 이는 보고**의 미러 사이트 중 하나였고 86%의 유사도를 보였다. 이 사이트를 분석해 본 결과 재오픈할 때 웹페이지 소스 일부분을 리뉴얼 하면서 사용하지 않는 부분에 주석 처리를 많이 하였기 때문인 것으로 나타났다.

4.4 외국 상위 10개 해적 사이트 실험 결과

2015년 1월 4일 Alexa 기준 외국 상위 10개 해적 사이트의 미러사이트를 동일한 방법으로 찾아보았다[9]. 이를 중 2015년 10월 15일 기준으로 하나의 사이트는 서비스를 종료하였기에 상위 9개 해적 사이트를 대상으로 실험하였다. 도메인을 수집한 결과 805개의 도메인이 검색되었고, 총 25개의 해적 사이트가 있었다. 이 도메인들을 대상으로 LCS 알고리즘을 적용한 결과를 표 4에 요약하였다.

표 4 LCS 유사도 측정 결과(외국)

Table 4 Result of LCS similarity measurement (international)

	Threshold (%)					
	50	60	70	80	85	90
True Positive	23	23	23	23	22	22
False Positive	1	0	0	0	0	0
False Negative	2	2	2	2	3	3

외국 해적 사이트의 경우 역시 탐지율 자체는 높았으나 미탐이 2개가 존재했다. 이 경우는 해당 해적 사이트의 태그 개수는 44개밖에 되지 않았지만 두 미러 사이트의 경우 태그 개수가 95개로 2배 이상 많았으며, 이 두 미러 사이트와의 LCS 값은 44로 모두 같았지만, 나머지 51개 태그 모두가 <a> 태그로 검색어 추천으로 인한 링크였다.

4.5 트위터를 이용한 추가 도메인 검사

제안된 연구는 검색에 노출된 도메인만을 가지고 하였기 때문에 도메인이 노출되지 않으면 검색되지 않는다는 단점이 있다. 야**, 졸라**, 토판**, 에스** 4개의 검색되지 않는 도메인을 대상으로 조사를 한 결과 에스**를 제외한 3개의 사이트의 경우 트위터를 통하여 도

메인 주소의 링크를 공유하고 있었다. 또한 검색에 직접적으로 노출되었던 국내 도메인 중 13개의 도메인이 트위터를 통해서도 주소를 공유하였다. 반면 외국의 경우에는 단 한 곳만이 트위터를 이용하여 추가적인 공유를 하고 있었다. 이러한 이유는 국내의 경우 도메인이 신고에 의해 쉽게 차단당하기 때문에 빠르게 사이트를 활성화시키기 위한 것으로 보인다.

5. 결 론

우리는 불법 사이트들이 도메인만을 변경하여 재오픈하였을 때 자동으로 탐지하는 기술에 대해 제안하였다. 구글 검색과 LCS 유사도 측정 알고리즘을 이용하였고, 실험 결과 대부분의 불법 사이트들이 사이트를 재오픈할 때 기존에 사용하던 웹페이지 소스 코드를 그대로 사용한다는 것을 알 수 있었다. 이러한 특징들을 봤을 때 제안된 시스템을 이용한다면 불법 사이트들이 재오픈 하더라도 다시 차단하기 쉬울 것으로 예상된다.

References

- [1] Business Software Alliance, "Shadow Market: 2011 BSA Global Software Piracy Study," May 2012.
- [2] A. Adermon, C-Y. Liang, "Piracy and Music Sales: The Effects of An Anti-Piracy Law," *Journal of Economic Behavior & Organization*, Vol. 105, pp. 90-106, Sep. 2014.
- [3] B. Danaher, MD. Smith, R. Telang, "The Effect of Piracy Website Blocking on Consumer Behavior," *Social Science Research Network*, Nov. 2015.
- [4] KOCSC. (2016, Jan. 6). [Online]. Available: http://www.kocsc.or.kr/02_infoCenter/info_Communication_List.php
- [5] A. P. E. Rosiello, E. Kirda, C. Kruegel, "A Layout-Similarity-Based Approach for Detecting Phishing Pages," *SecureComm*, pp. 454-463, Sep. 2007.
- [6] S. Roopak, T. Thomas, "A Novel Phishing Page Detection Mechanism Using HTML Source Code Comparison and Cosine Similarity," *Advances in Computing and Communications (ICACC)*, pp. 167-170, Aug. 2014.
- [7] M. E. Maurer, D. Herzner, "Using visual website similarity for phishing detection and reporting," *CHI'12 Extended Abstracts on Human factors in Computing systems*, pp. 1625-1630, May. 2012.
- [8] S. Brin, L. Page, "Reprint of: The Anatomy of a Large-Scale Hypertextual Web Search Engine," *Computer networks, 2012 Fourth International Conference on*, Vol. 56, No. 18, pp. 3825-3833, Oct. 2012.
- [9] Torrentfreak. (2015, Jan. 04). [Online]. Available: <https://torrentfreak.com/top-popular-torrent-sites-2015-150104/>



이기룡

2014년 경북대학교 컴퓨터공학과 졸업
(학사). 2014년~현재 고려대학교 컴퓨
터·전파통신공학과 석사과정. 관심분야는
네트워크 보안



이희조

1989년~2000년 포항공대 컴퓨터공학과
학사/석사/박사. 2000년~2001년 미국
Purdue University 박사후연구원. 2001
년~2003년 안철수연구소 최고기술책임
자(CTO). 2004년~현재 고려대학교 컴
퓨터학과 교수. 관심분야는 네트워크 보
안, 인터넷웹/DDoS 공격 대응기술, 고가용성 시스템 설계