트래픽 모니터링을 통한 P2P 및 웹 하드 다운로드 응용의 파일이름 식별 방법

(A File Name Identification Method for P2P and Web Hard Applications through Traffic Monitoring)

손 현 구 † 김 기 수 †† 이 영 석 †††
(Hyeongu Son) (Kisu Kim) (Youngseok Lee)

요 약 최근 파일 공유 및 인터넷 전화, 동영상 스트리밍 같은 진화된 인터넷 응용 서비스들이 등장하고 있다. 특히 P2P 또는 웹 기반 파일 공유 응용 들은 컨텐츠 불법 복제와 소수 사용자에 의한 다량의 트래픽 점유율 등의 문제를 지속적으로 제기하고 있다. 본 논문에서는 트래픽 모니터링을 통하여 P2P 응용 및 웹하드 응용에서 다운로드 받는 파일이름을 식별하는 방법을 제안하고 이의 실험 결과를 제시한다. 파일 이름을 식별하기 위해서 패킷 페이로드 내에 존재하는 한글 문자열을 디코딩하는 방법을 이용하였고, BitTorrent, 클럽박스 및 tple을 대상으로 실험하여 다운로드받는 파일이름을 탐지할 수 있음을 보였다.

키워드 : P2P, 웹하드, 트래픽 모니터링, 파일이름, 식별, 한글

Abstract Recently, advanced Internet applications such as Internet telephone, multimedia streaming, and file sharing have appeared. Especially, P2P or web-based file sharing applications have been notorious for their illegal usage of contents and massive traffic consumption by a few users. This paper presents a novel method to identify the P2P or web-based file names with traffic monitoring. For this purpose, we have utilized the Korean decoding method on the IP packet payload. From experiments, we have shown that the file names requested by BitTorrent, Clubbox, and Tple applications could be correctly identified.

Key words : P2P, webhard traffic monitoring, file name, identification, Korean

1. 서 론

WWW(World Wide Web)의 등장은 인터넷이 널리 보급될 수 있는 계기를 만들어 주었다. 2000년 이전까지는 인터넷 사용자는 단순 웹 검색, 메일 전송 및 채팅 등의 기능만을 사용했다. 하지만 2000년 이후 네트워크의 성능이 빠르게 향상되면서 인터넷을 이용한 전화, 동영상 스트리밍 및 파일 공유 등의 다양한 서비스들이 등장하고 있다. 특히 P2P를 이용한 파일 공유는 1999년 6월 미국에서 냅스터라는 음악 파일 공유 서비스가 출현한 후 널리 사용되고 있으며, 최근에는 BitTorrent, 프루나 등의 다양한 P2P 응용 프로그램이 사용되고 있다.

P2P는 사용자가 가진 정보 또는 파일을 쉽게 공유할 수 있는 장점이 있다. 하지만 P2P 사용자는 하나의 파일을 다운로드 할 때 파일을 작은 크기 단위로 나누어 여러 사용자들로부터 다운로드하기 때문에 네트워크에 많은 양의 트래픽을 발생시킨다. 이런 네트워크 트래픽 양의 증가는 네트워크 성능의 저하시킨다. 또한 네트워크의 성능 저하는 단순히 웹 검색 등만을 사용하는 일반 인터넷 사용자에게 피해를 줄 수 있다. 또한, P2P를 이용하여 받은 파일에서 악성 코드가 발견되기도 한다.

· 본 연구는 지식경제부 및 한국산업기술평가관리원의 산업원천기술개발사업(정보통신)의 일환으로 수행하였음[KI001878, 초정밀 측정 및 분석 기술 연구]

† 학생회원 : 충남대학교 컴퓨터공학과
hgson@cnu.ac.kr
†† 비 회 원 : 충남대학교 컴퓨터공학과
berserk555@naver.com
††† 정 회 원 : 충남대학교 컴퓨터공학과 교수
lee@cnu.ac.kr
(Corresponding author임)

논문접수 : 2010년 5월 31일
심사완료 : 2010년 8월 18일

정보과학회논문지 : 정보통신 제37권 제6호(2010.12)

이 악성 코드는 자신의 컴퓨터에 감염되어 자신의 파일 또는 정보가 유출되거나 다른 서버/컴퓨터를 공격하는 좀비 PC로 활동할 가능성이 높다. 따라서 네트워크 성능 제어 및 보안을 위하여 P2P 트래픽을 탐지하고 제어할 필요가 있다.

P2P 트래픽이 네트워크에 과부하를 발생시키기 때문에 학교, 공공기관 등에서는 이들 트래픽을 생성시키는 프로그램의 사용을 제한하기 위하여 방화벽, 라우터 및 백본 스위치에서 잘 알려진 P2P 응용 프로그램이 사용하는 포트번호를 사용한다. 하지만, 최근 등장한 P2P 응용 프로그램들은 80번 포트번호를 이용하여 파일을 검색하고, 검색된 파일을 다운로드 받을 때 동적 포트번호를 사용한다. 방화벽에서는 이런 동적 포트번호를 제한하여 P2P 응용 서비스들을 제한하고 있기 때문에, 최근 출현하는 P2P 응용 서비스들은 별도의 제한을 받지 않는 HTTP(Hyper-Text Transfer Protocol) 기반의 80번 포트번호를 사용한다.

HTTP는 P2P 및 웹 하드 응용에서 사용자가 파일 검색을 위하여 파일 이름 저장 서버에 검색을 요청하고 응답을 받을 때 사용된다. 따라서 사용자가 파일 검색을 요청할 때 전송하는 HTTP 요청 패킷과 응답 패킷을 디코딩하여 다운로드 할 파일 이름의 추출이 가능하다. 하지만 많은 HTTP 패킷들이 gzip과 같은 압축을 통해 페이로드를 압축하고, 알파벳이 아닌 경우 다국어 인코딩을 사용한다. 트래픽 모니터링에서 이런 HTTP 패킷에서 압축 해제 및 다국어 디코딩 방법을 적용한 사례가 없었다. HTTP 패킷에 적용된 압축해제는 HTTP 헤더에 명시된 압축 방법에 따라 수행된다. 또한 다국어 인코딩이 적용된 HTTP 패킷의 경우 각 나라에서 사용되는 인코딩 정보에 따라 디코딩을 통해 파일 이름을 식별해야 한다. 따라서, 본 논문에서는 HTTP를 기반으로 한 P2P 및 웹하드 응용에서 사용되는 트래픽으로부터 파일 이름 식별 방법을 제시한다. 특히, 국내에서 다운로드되는 파일의 이름 대부분이 한글로 되어 있기 때문에 이를 디코딩하기 위하여 [1]에서 제안된 한글 디코딩 방법을 이용한다.

대부분의 P2P 응용은 대부분 각자의 프로토콜을 정의하여 이용하기 때문에 모든 P2P 응용에 대하여 파일 이름을 식별할 수 있는 일반적인 방법을 개발하기가 쉽지 않다. 따라서, 본 논문에서는 BitTorrent[7], 클럽박스[8], Tple[9]이라는 대표적인 세 가지 P2P 및 웹 하드 응용을 대상으로 트래픽 모니터링을 통하여 사용자가 요청하는 파일 이름 및 다운로드 파일의 크기 등의 정보를 식별한다.

본 논문의 구성은 2장에 관련연구, 3장에 P2P 파일 이름 식별 방법을 제시한다. 실험환경과 실험결과는 각각 4장과 5장에 기술하며, 결론 및 향후 연구는 6장에 기술한다.

2. 관련연구

[2]는 일본 내에서 2005년과 2008년 인터넷 사용률 변화를 제시하고 있다. [2]에 의하면 TCP를 사용하는 80 포트의 사용률이 증가함을 알 수 있으며, 이에 비해 동적으로 사용되는 포트번호의 사용은 둔화됨을 알 수 있다. 이는 동적 포트를 사용하는 P2P 응용 프로그램이 80 포트와 같은 잘 알려진 포트번호를 이용하기 때문이다. 따라서 잘 알려진 포트번호를 이용하는 P2P 트래픽의 탐지가 필요하다.

[3]에서의 P2P 분류 방법은 잘 알려진 포트번호를 이용한다. 하지만 포트번호만을 이용할 경우 P2P 트래픽을 분류하기 어렵다. 왜냐하면, 최근 등장하는 P2P 트래픽은 잘 알려진 포트번호 또는 동적인 포트번호를 사용하기 때문에 이를 이용하면 P2P 트래픽 분류의 결과가 부정확해지는 단점이 있다. 이런 단점을 보완하기 위하여 시그너쳐 기반 P2P 트래픽 탐지 방법이 제시되었다[4]. [4]와 [5]는 P2P 트래픽을 탐지할 수 있지만 P2P 트래픽이 어떤 용도로 사용되었으며, 어떤 파일을 사용자가 받았는지 알 수 없다.

네트워크상에서 패킷을 캡쳐하고 패킷의 내용을 보기 위한 툴로써 wireshark[6]이 널리 사용되고 있다. wireshark는 캡쳐된 패킷의 헤더 내용을 보여주지만, 디코딩된 정보가 알파벳을 기반으로 되어 있어 한글 등 알파벳을 사용하지 않을 경우 이에 대한 결과를 확인하기는 쉽지 않다. 왜냐하면, 한글의 경우 한 글자마다 초성, 중성, 종성의 인코딩 정보를 보고 디코딩을 해야 하기 때문이다. 이러한 패킷들이 다수 있을 경우 이를 디코딩하여 트래픽을 분석하는 것은 많은 시간을 소요하게 한다. 따라서 본 논문에서는 [1]에서 제시된 한글 디코딩 방법을 이용하여 P2P 및 웹하드 트래픽에서 다운로드 파일 이름 식별 방법을 본 논문에서 제안한다.

3. P2P/웹 하드 파일이름 탐지 방법

P2P 및 웹 하드 서비스로부터 다운로드 되는 파일 이름을 식별하기 위하여 현재 많이 사용되는 P2P 응용 프로그램인 BitTorrent와 대표적인 웹 하드 서비스인 클럽박스와 Tple를 이용한 파일 이름 탐지 방법을 제시한다.

3.1 BitTorrent 파일 이름 탐지 방법

BitTorrent에서 사용자의 파일 다운로드는 확장자가 torrent라는 메타 파일을 다운로드 한 후 BitTorrent에서 제공하는 툴을 이용하여 실제 파일을 다운로드 한다. 이때 사용자가 다운로드 받은 메타 파일은 사용자가 다운로드 할 파일의 정보를 포함한다.

사용자가 BitTorrent를 이용하여 실제 파일을 다운로

표 1 torrent 메타 파일의 구조

info	length	다운로드 파일의 전체 크기
	name	다운로드 파일이름
	pieces	파일조각에 대한 해쉬 키 묶음 문자열
	pieces length	파일조각 길이
announce	Tracker의 URL	

그림 2 torrent 메타 파일에서 다운로드할 파일의 정보

드하기 위하여 받은 메타파일의 구조는 표 1과 같다. 메타 파일은 info와 announce로 구성되며, info에 실제 다운로드할 파일의 이름과 전체 크기 및 파일의 조각에 대한 길이와 파일의 무결성을 확인하기 위한 해쉬키가 존재한다. announce에는 P2P 중계 서버 역할을 하는 tracker의 URL이 포함된다. 사용자는 이 메타 파일을 이용하여 다운로드 할 파일의 seeder와 peer에 대한 정보를 가져온다.

BitTorrent에서 다운로드할 파일 이름을 탐지하기 위하여 본 논문에서는 사용자가 웹 서버에 게시된 메타 파일을 다운로드 받을 때 HTTP 응답 메시지를 분석한다. 그림 1은 사용자가 받은 메타파일의 HTTP 응답 메시지의 HTTP 헤더의 내용이며, 그림 2는 HTTP 응답 메시지에 포함된 메타 파일의 정보를 보여준다. 사용자가 다운로드한 메타 파일의 파일 이름은 그림 1처럼 HTTP 응답 패킷의 HTTP 헤더로부터 식별한다. 메타 파일의 이름은 그림 1에서 박스안의 내용처럼 "Filename=" 으로 시작하며, 그림 1에서 제시된 메타 파일 이름은 "[T] 추노 - E11 - 100210 28SD29.torrent"이다.

사용자가 다운로드할 파일의 정보는 그림 2에서 제시된 메타 파일에 포함되어 있다. 메타 파일에는 다운로드할 파일의 이름 및 크기와 파일 이름에서 한글이 인코딩된 타입이 명시되어 있다. 파일의 개수가 여러 개 일

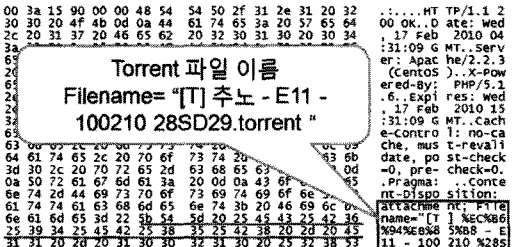

그림 1 torrent 메타 파일을 다운로드 할 때 HTTP 패킷에서 메타 파일의 이름

경우는 디렉토리로 묶어 사용자가 다운로드하는 전체 파일 이름의 확인이 가능하다.

그림 1과 그림 2에서 제시된 것처럼 파일 이름을 식별하기 위한 과정은 2단계로 구성된다.

- 1단계 : BitTorrent 메타 파일 다운로드 인지 단계
- 2단계 : 다운받은 메타 파일로부터 실제 다운로드할 파일의 정보를 식별하는 단계

1단계에서는 HTTP 헤더에서 torrent 메타 파일 이름의 식별을 통해 사용자가 메타 파일 다운로드를 인지한다. 메타 파일 이름의 식별은 HTTP 헤더에서 "Filename"이라는 시그너쳐를 찾은 다음 파일의 확장자가 ".torrent"인지를 확인한다. torrent 메타 파일이 다운로드 된 것이 인지되면 2단계를 수행하여 실제 다운로드할 파일의 정보를 식별한다.

실제 다운로드 파일 정보의 식별을 위하여 3가지 시그너쳐를 사용한다. 한글 인코딩 방법은 'encoding', 파일 길이는 'length', 다운로드 파일 이름은 'name'을 시그너쳐로 사용한다. Torrent의 메타 파일은 표 2와 같은 방법으로 작성된다. 따라서 표 2에서 제시된 방법을 통해 다운로드할 파일의 이름과 크기를 식별한다.

표 2 메타파일에 인코딩된 데이터를 추출하는 방법

```
byte string - 4:spam--->"spam"
integers - i3e--->"3"
lists - l4:spam4:eggse--->"spam","eggs"
 dictionaries-
d3:cow3:moo4:spame--->"cow","moo","spam"
```

3.2 클럽박스 파일 이름 탐지 방법

클럽박스는 국내에서 많이 사용되는 웹 하드 응용 서비스 중 하나이다. 사용자가 클럽박스에서 파일 다운로드를 위하여 다운로드 관리 프로그램이 실행되고, 파일을 다운로드 받을 때는 일반적인 P2P 응용처럼 피어와 연결되어 파일을 다운로드한다.

클럽박스로부터 파일을 다운로드 하기 전 그림 3과 같은 다운로드 파일에 대한 결제를 위한 팝업 창이 실

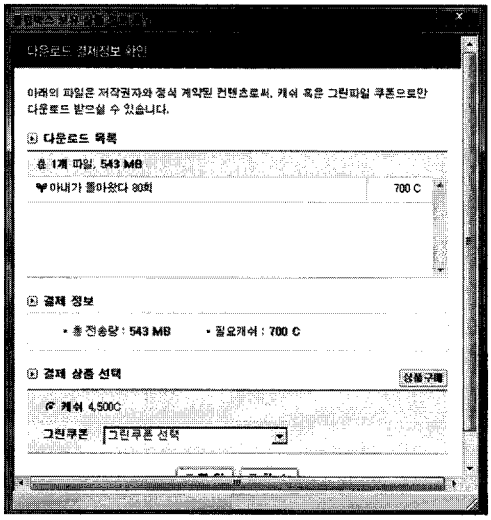

그림 3 클럽박스 사용자가 파일 다운로드하기 전 결제를 위한 팝업 창

표 3 클럽박스 파일 다운로드 창 내용

<meta http-equiv="Content-Type" content="text/html; charset=euc-kr" />
<meta name="keywords" content="클럽박스, clubbox, 자료실" />
<title>클럽박스 보유 상품 정보 확인</title>

행된다. 그림 3으로부터 사용자가 다운로드 하는 파일의 개수, 전체 파일의 크기 뿐만 아니라 사용자의 캐쉬 정보 및 다운로드할 파일의 캐쉬 정보 등의 확인이 가능하다. 본 논문에서는 그림 3과 같은 팝업 창이 로딩될 때 사용되는 패킷을 분석하여 클럽박스에서 사용자가 다운로드 받는 파일의 이름을 식별한다.

표 3은 그림 3의 팝업 창이 로딩될 때 HTTP 패킷에서 나타나는 HTML 태그의 일부분을 보여준다. 표 3에 제시된 내용에서 한글로 된 파일 이름이 인코딩된 타입 정보(charset=euc-kr)를 얻을 수 있다.

그림 4는 그림 3의 팝업 창이 로딩될 때 HTML 태그 중에서 다운로드할 파일의 정보를 보여준다. 그림 4의 상단에 박스모양으로 된 부분이 사용자가 다운로드할 파일의 개수와 총 용량을 나타내며, 이를 분석하기 위한 시그너쳐로 'total_file'을 사용한다. 또한, 다운로드할 파일 이름과 다운로드 시 필요한 캐쉬의 정보도 알 수 있다.

3.3 Tple 파일 이름 탐지 방법

클럽박스외에 국내 웹 하드 응용의 하나인 Tple에서 다운로드 되는 파일 이름을 식별하는 방법은 다음과 같다. Tple도 클럽박스와 같이 실제 다운로드 하기 전에

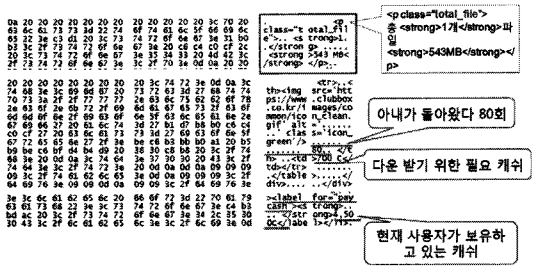

그림 4 클럽박스 파일 탐지 예시

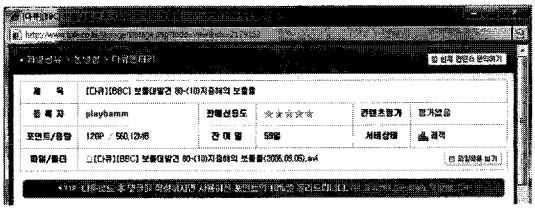

그림 5 Tple에서 파일 다운로드를 위한 팝업 창

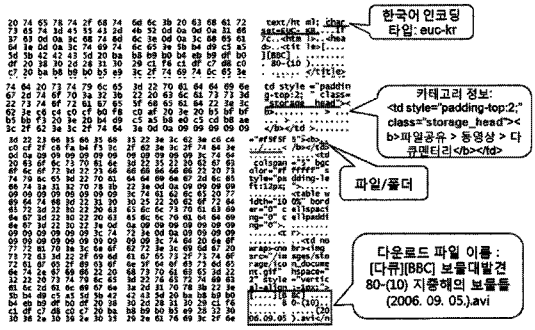

그림 6 Tple에서 다운로드 파일 식별 예시

다운로드할 파일의 정보가 보이는 팝업 창이 로딩된다. 이때 사용자가 다운로드할 파일의 이름과 크기가 표시된다. 또한 Tple에서는 상위에 다운로드할 파일이 Tple에서 어느 카테고리에 포함되어 있는지까지 나타나 있다. 따라서 Tple에서 다운로드할 파일의 이름과 크기의 식별뿐만 아니라 해당 파일의 카테고리 정보까지 알 수 있다.

트래픽 모니터링을 통하여 다운로드할 파일의 이름을 식별하기 위하여 그림 6처럼 HTTP 패킷에 포함된 HTTP 헤더와 HTML 태그를 분석한다. Tple은 클럽박스와는 다르게 한글 인코딩 정보가 HTML 태그에 포함되는 것이 아니라 HTTP 헤더의 'charset'의 필드에 제공되기 때문이다.

HTML 태그에서 다운로드할 파일의 이름과 크기 및 카테고리 정보를 추출하기 위하여 그림 6의 시그너쳐를 사용한다. 카테고리 정보는 HTML 태그에서 'storage_head', 파일 용량은 '포인트', 파일 이름은 '파일/폴더'를 이용한다.

4. 실 험

4.1 실험환경

본 논문에서는 그림 7처럼 충남대 네트워크와 인터넷 사이의 라우터에서 tcpdump[10]를 이용하여 트래픽을 수집하였다. 수집된 트래픽을 분석하여 bitTorrent와 클럽박스, Tple을 통하여 사용자가 다운로드 받는 파일의 정보를 식별하였다. 3가지 P2P/웹하드 응용으로부터 다운로드하는 파일 이름 식별 실험을 위하여 본 논문에서는 각 응용별로 100번씩 파일 다운로드를 시도하였다.

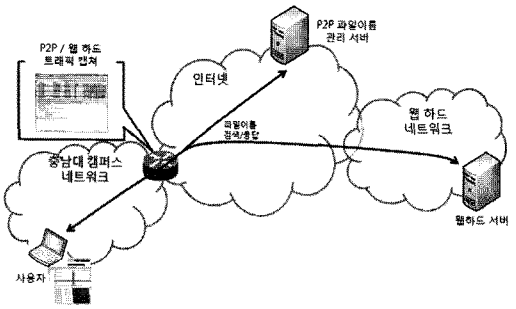

그림 7 P2P 및 웹 하드 다운로드 파일 이름 식별을 위한 트래픽 모니터링 환경

4.2 실험결과

그림 7의 실험환경에서 수집한 트래픽에서 본 논문에서 제시한 다운로드 파일 이름 식별 방법을 이용하여 그림 8과 같은 결과를 얻었다. 그림 8은 bitTorrent의 다운로드 파일 이름과 크기 식별 결과 및 사용자의 IP 주소와 포트번호를 보여준다. 클럽박스와 tple의 경우 파일 다운로드시 별도의 메타파일을 다운로드 하지 않으므로 두 웹하드 응용에서 파일 이름 식별결과에는 메타 파일이름 식별결과는 포함되지 않는다.

다운로드 파일 이름 식별의 정확도를 실험하기 위하여 본 논문에서는 각 응용별로 100번씩 bitTorrent 메타파일 다운로드를 수행하고, 웹 하드에서는 파일 다운로드를 시도하였다. 그림 7의 환경에서 캡쳐된 트래픽을 분석하여 그림 8과 같이 다운로드 파일 이름을 식별하고 정확도를 분석하였다.

Session Information 1			
src.IP	dst. IP	src. port	dst. port
208.93.233.153	168.188.46.122	80	52028
Torrent File Name	"[T] 추노 - E11 - 100210 28SD29.torrent"		
Download File Name	추노.E11.100210.HDTV.450p.H264-Angel.avi		
Download File Size	758.565MB		

그림 8 bitTorrent에서 다운로드 파일 이름 식별 결과 예제

표 4는 3가지 P2P/웹 하드에 대한 실험 결과를 보여준다. 본 논문에서 제시한 3가지 P2P 웹 하드 응용에서 다운로드 파일 이름 식별 방법을 이용하여 실험한 결과 100% 모두 탐지하였다. BitTorrent와는 다르게 클럽박스와 tple의 경우 파일 다운로드 할 때 메타 파일을 별도로 다운로드 하지 않으므로 메타 파일 이름 식별결과는 존재하지 않는다.

6. 결론 및 향후 과제

본 논문에서는 트래픽 모니터링을 통하여 P2P 및 웹 하드 응용 사용자가 다운로드 받을 파일의 이름을 식별할 수 있음을 보였다. 이는 포트번호별 또는 시그너처를 이용하여 P2P 트래픽 탐지 단계를 벗어나 사용자가 어떤 파일을 어떻게 다운로드 했는지 알 수 있다.

최근 인터넷 상에서 영화, 음악 등의 지적재산권과 관련있는 컨탠츠의 불법 다운로드가 증가하고 있다. 사용자가 이들 파일을 다운로드 받고자 할 때 본 논문에서 제시한 파일 이름 식별 방법을 적용한다면 여러 컨탠츠의 지적재산권 침해를 예방할 수 있으며, 이들 트래픽을 지속적으로 감시하는 것이 가능하다. 또한 기업에서 기밀로 다루고 있는 것들에 대해 제안한 파일 이름 식별 방법을 통해 기업망을 감시한다면, 해당 기업의 기밀 유출 방지에도 활용이 가능하다.

본 연구에서 파일 이름 식별을 위하여 실시간 트래픽 모니터링이 아닌 이미 캡쳐된 트래픽을 이용하였다. 1~10Gbps와 같은 고속의 링크에서 본 논문에서 제안한 파일 이름 식별 방법을 적용할 수 있는 연구가 필요하다. 하지만, 고속의 링크에서 파일 이름 식별과 같은 DPI (Deep Packet Inspection)는 쉽지 않기 때문에 캡쳐된 패킷을 병렬처리 또는 분산처리 등을 이용하면 파일 이름 식별 속도가 향상될 것으로 예상된다.

향후 본 연구를 확장하여 한글이 아닌 일본어 및 중국어를 사용하는 P2P 및 웹 하드 트래픽으로부터 다운로드 파일 이름을 식별하는 방법의 제시가 필요하다.

표 4 P2P 웹 하드 다운로드 파일 이름 식별 결과

P2P/웹하드 응용	파일 다운로드 회수	다운로드 파일 이름 탐지율(%)	메타파일 이름 탐지율(%)	파일 용량 탐지율(%)
bitTorrent	100	100	100	100
클럽박스	100	100	-	100
tple	100	100	-	100

참 고 문 헌

[1] Kisu Kim, Hyeongu Son, Taeck-geun Kwon, Youngseok Lee, "A Korean Decoding Method for Content Classification of HTTP Traffic," KICS Fall Conference, Nov. 2008.
[2] K. Cho, K. Fukuda, H. Esaki and A. Kato, "Observing Slow Crustal Movement in Residential User Traffic," ACM CoNext, no.12, Dec. 2008.
[3] Packetteer, http://www.packeteer.com/.
[4] S. Sen and J. Wang, "Analyzing Peer-to-Peer Traffic Across Large Networks," *IEEE/ACM Transactions on Networking*, vol.12, no.2, pp.219-232, April 2004.
[5] S. Sen, O. Spatscheck, and D. Wang, "Accurate, Scalable In-Network Identification of P2P Traffic Using Application ignatures," *ACM WWW*, pp. 512-521, May 2004.
[6] Wireshark, http://www.wireshark.org/.
[7] BitTorrent Protocol, http://www.bittorrent.com/.
[8] Clubbox, http://www.clubbox.co.kr/.
[9] Tple, http://www.tple.co.kr/.
[10] tcpdump, http://www.tcpdump.org/.

손 현 구
정보과학회논문지 : 정보통신
제 37 권 제 4 호 참조

김 기 수
2008년 충남대학교 컴퓨터공학과 학사.
2010년 충남대학교 컴퓨터공학과 석사.
관심분야는 인터넷 트래픽 분석 등

이 영 석
정보과학회논문지 : 정보통신
제 37 권 제 4 호 참조