웹 하드 디스크 응용 트래픽 분석

배준호, 이현신, 윤성호, 김명섭
고려대학교 컴퓨터정보학과
e-mail : {krjj21, oshyuns, hiption, tmskim}@korea.ac.kr

Web Hard Disk Application Traffic Analysis

Joon-Ho Bae, Hyun-Shin Lee, Sung-Ho Yoon, Myung-Sup Kim
Dept. of Computer and Information Science, Korea University

요 약

최근 인터넷 사용량이 급증함으로 인해 인터넷 응용프로그램의 개발, 종류 및 사용이 기하급수적으로 늘어나고 있다. 초기에 사용되었던 웹, 파일전송, E-mail 등 well-known port 기반 서비스가 변화되어 unwell-known port 기반 서비스가 주를 이루고 있다. 이러한 상황에서 효율적인 네트워크 관리를 위한 응용 트래픽의 분류가 어려워지고 있으며, 효과적인 트래픽 분류를 위한 연구가 많이 진행되고 있다. 본 논문에서는 가장 많이 사용되는 인터넷 응용프로그램인 웹하드를 대상으로 이들로부터 발생하는 트래픽을 분석하였다. 먼저 웹하드의 정의 및 기능과 그 종류에 대해서 기술하였다. 또한, signature 기반 분류 방법으로 웹하드 트래픽을 패킷 캡쳐 툴과 Flow 을 이용하여 분석하고 그 결과로부터 응용 트래픽 분류의 관점에서의 웹하드 트래픽의 특징에 대해서 기술하였다.

서론

오늘날에는 대용량 초고속 광통신 인터넷 서비스의 구축이 가능해짐에 따라 고용량, 고품질의 미디어 매체를 접할 수 있게 되었다. 이러한 미디어 매체를 제공하는 서비스로는 P2P 서비스와 웹하드 서비스가 있다. 두 서비스 중, 본 논문에서는 웹하드 서비스를 중점적으로 분석하였다.

본 논문에서는 많은 종류의 웹하드 서비스와 업체의 증가로 웹하드 서비스에 대한 중점적인 분석이 필요함을 느끼고, 국내 웹하드 서비스 중 61 개를 선별하여 발생하는 트래픽의 특징을 분석 및 정리하였다.

분석을 위한 1 차적인 방법으로는 수작업으로 분석이 필요하다. 패킷 캡쳐 툴을 사용해 웹하드 클라이언트 프로세스에서 발생하는 패킷을 캡쳐한 후 header signature(Ex IP, Port)를 찾고 정리하여 XML 파일로 정형화 했으며, 2 차적으로 검증 시스템을 통하여 header signature payload signature 를 재정리하고, payload signature 를 추가하여 웹하드 트래픽의 특징 분석을 위한 웹하드 트래픽 분류 신뢰성을 높였다.

본 논문의 각 장의 구성은 다음과 같다. 2 장에서는 응용 트래픽 분류에 관련된 연구에 대해서 자세히 설명하고, 3 장에서는 패킷 캡쳐 툴을 통한 분류 방법 및 분류 검증 시스템을 통한 signature 생성 방법을 설명한다. 4 장에서는 분류 기준 및 방법을 설명한다. 5 장에서는 검증 방법에 대한 설명을 하고, 6 장에서는 실험 방법을 설명하고, 실험의 결과에 따라서 웹하드 서비스를 분류 정리한다. 마지막으로 결론 및 향후 과제를 언급하며 본 논문을 마친다.

2. 관련연구

2.1 웹하드 디스크 응용

웹하드 또는 웹하드 디스크란 인터넷 상에 일정용량의 공간(RAID, SCSI HDD 등)을 제공하며 DVD, CD, USB 메모리 등의 운반 가능한 저장매체 없이도 인터넷이 연결된 곳이면 어느 곳에서나 파일 등을 저장, 편집, 실행하고 다수의 사람들과 공유할 수 있게 해주는 가상 하드디스크이다.[1] 웹하드는 기본적으로 파일의 검색, 다운로드, 업로드의 서비스를 제공한다. 국외에서는 웹하드보다 P2P 의 사용이 많으며, 웹하드의 쓰임도 국내와는 다르게 자신의 저장공간을 마련해두고 언제 어디서든지 사용할 수 있게 개인용도의 서비스로만 사용된다. 하지만 국내에서는 보통 유료화 되어있음에도 불구하고 P2P 와 달리 빠른 속도의 다운로드 및 업로드와 안정성 등의 장점으로 활발하게 서비스되고 있다. 국내에서 서비스되고 있는 웹하드 서비스는 60 여 개가 넘으며, 업체별로 클라이언트 프로그램과 커뮤니티를 제공하고 있고, 각기 다른 프로토콜을 사용하고 있다.[2] <표 1>의 자료는 웹하드 서비스들을 동작 방식에 따라 분류하고 한 것이다

2.2 Signature 기반 분류 방법

Signature 란, 패킷의 헤더정보나 페이로드에서 발견되는 일정한 패턴으로, 특정한 응용에 의해서 발생된 플로우나 트래픽을 해당 응용에 속한것으로 판정될 만큼 빈번이 발견되어야 한다.[3]

이 논문은 2007 년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임(KRF-2007-331-D00387)

<표 1> 동작 방식에 의한 웹하드 분류

동작방식	설명	웹하드 서비스명
웹하드	회사에서 대형 저장공간을 제공하고 사용자들이 일부분을 임대하여 사용하는 형태이고, 독립된 응용프로그램 사용.	100 기가, 넷폴더, 다이하드, 디스크팝, 몽디스크, 싼타 25, 썬폴더, 엑스톡, 엔디스크, 쿨디스크, 큐파일, 크레이지파일, 토마토팡, 토토디스크, 파일콘, 디스크펌프, 폴더플러스, 하드스토어, 쏘디스크, OTL 파일
클럽형 웹하드	웹하드 방식과 비슷하며 클럽 관리자에게 모든 법적인 책임을 부여하고 클럽회원에게만 자료를 공유	다운즈, 더파일, 디노파일, 디스크스토리, 따오기, 로또파일, 메가파일, 모모디스크, 바다바, 보물박스, 슈퍼파일, 심파일, 아이팝, 엑스파일, 오케이공유, 와와디스크, 요타디스크, 짱클럽, 클럽하드, 토마토파일, 파일몬, 프리팝, 피디팝, 브이하드
웹방식 웹하드	웹하드 방식과 비슷하지만 사용자들이 회사의 서버에 자료를 올리면 해당회사에서의 웹 게시판을 통해서 사용자들이 다운로드 하는 방식	7 디스크, 뉴와우디스크, 아삼박스, 엠파일, 온디스크, 위디스크, 제트파일, 짱파일, 캐치파일, 케이디스크, 코코디스크, 클럽박스, 파일시티, 피디박스, 피코팟, AM

패킷의 헤더정보에 의해 만들어진 signature 를 header signature 라고 하며, 보통 ip address, port, protocol 등이 주를 이룬다. payload signature 는 header signature 와 달리 payload 에서 발견된 패턴을 말한다. 본 논문에서는 이 두 개념을 이용한 분류 방법을 signature 분류방법이라 정한다.

3. 웹하드 응용 시그니쳐 생성 방법

3.1 SocketSniff 를 이용한 signature 생성

SocketSniff[4]는 선택한 프로세스에서 나가는 패킷을 캡쳐하는 패킷 캡쳐 툴로써, 특정 프로세스의 정보(socket number, protocol, local-address, local-port, remote-address, remote-port) 등을 알 수 있다. 각각의 웹하드 서비스를 사용하면서 발생되는 패킷을 캡쳐하였다. 패킷을 캡쳐 한 후에 header signature 를 추출하기 위해 엑셀문서로 정리하고 특정 패턴을 찾아내어 XML 파일을 작성시 참고 하였다. 다음은 "파일시티" 웹하드 서비스의 패킷들의 예이다.

<표 2> "파일시티" 웹하드 패킷들

Name	Protocol	Remote IP	Remote port
…	…	…	…
Filecity	TCP	xxx.xxx.230.29	59781
Filecity	TCP	xxx.xxx.230.29	54094
Filecity	TCP	xxx.xxx.230.29	53801
Filecity	TCP	xxx.xxx.230.29	49187
Filecity	TCP	xxx.xxx.230.29	57646
…	…	…	…

<표 2>를 보면 파일시티 웹하드 응용은 파일전송을 위해 복수개의 연결을 맺지만 remote ip 에서 xxx.xxx.230.29 의 일정한 패턴을 보이고 있다. 이는 "파일시티"만이 가지는 고유한 패턴이라고 생각되므로 시그니쳐로 추출한다. payload signature 의 추출은 header signature 와 동일하며 payload 부분에 있는 패턴을 추출하게 된다.

<표 3> "짱파일" 플로우의 payload 데이터의 예

Destination IP address	Destination Port
xxx.xxx.232.20	2813
32 36 1D 00 00 00 00 00 D8 8A 02 00 FF FF 58 83 05	
32 36 1D 00 00 00 00 00 D8 8A 02 00 FF FF 58 83 05	
32 36 1D 00 00 00 00 00 D8 8A 02 00 FF FF 58 83 05	
32 36 1D 00 00 00 00 00 D8 8A 02 00 FF FF 58 83 05	
32 36 1D 00 00 00 00 00 D8 8A 02 00 FF FF 58 83 05	

<표 3>은 "짱파일"의 플로우 중 동일한 목적지를 가지는 payload data 의 예이다. <표 3>에서 패킷의 payload 부분에 일정한 패턴이 나타나므로 그 패턴을 payload signature 데이터로 간주한다. 이렇게 생성된 시그니쳐는 XML 파일의 형식으로 기술하고 이를 바탕으로 분석시스템이 동작한다. XML 파일의 특징은 데이터 관리 및 교환의 표준이므로 어느 프로그램에서나 데이터의 효율적 이용이 가능하고, 문서 내에 데이터를 계층적으로 표현하며 의미적으로 전달이 가능하다. (그림 1)은 분석 및 검증 시스템을 위해 사용한 XML 파일의 형식이다.

(그림 1) "짱파일"의 XML 파일

3.4 Flow 검증 시스템을 이용한 시그니쳐 추가 생성 및 재정리

수작업으로 인해 조사된 header signature 는 정확성이 높지 않다. 그러므로 체계적인 알고리즘을 이용한 Flow 검증 시스템을 통한 검증으로 header signature 의 오분류를 재정리하고 payload signature 를 필요에 따라 추가시켜서 웹하드 응용의 정확성을 높였다. 검증 시스템에 대해서 검증방법에서 자세히 다루도록 하겠다. 웹하드 서비스를 독립적으로 실행시킨 후 저자의 호스트들과 일치하는 Source IP address 를 기준으로 웹하드 프로세스의 이름을 비교하고, 출력된 결과의 FP(false positive), FN(false negative)을 중심으로 Flow 를 분석하는 테스트를 시행했다. 현재 이 시스템을 이용하여 피드백을 한 후, 분류의 90% 이상의 정확도로 웹하드 트래픽의 특징을 찾아 낼 수 있도록 하였다.

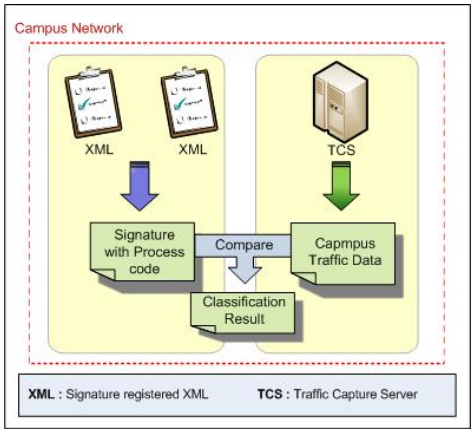
4. 분류 기준 및 방법

분류 방법은 학내 전체 트래픽에서 본 연구실에서 개발한 시스템으로 각 서비스의 서비스코드 및 프로세스를 비교하여 분류하였다. 효율적으로 웹하드 서비스를 분류하기 위해 다음과 같은 분류기준을 사용하였다.

<표 4> 분류기준

signature 생성 후 웹하드 서비스의 해당된 header signature 또는 payload signature 를 가진 플로우를 그 웹하드 서비스의 플로우로 분류한다.
동일한 header signature 를 가진 Internet Explore 의 웹 브라우징 서비스는 동일 웹하드 서비스로 분류한다.
동일한 header signature 를 가지고 payload signature 가 존재 하지 않은 어떤 웹하드 서비스는 같은 header signature 를 가진 다른 웹하드 서비스와 같은 그룹으로 묶는다.
웹하드 서비스에 포함된 각각의 프로세스들이 동일한 header signature 를 가지고 payload signature 가 존재하지 않는 경우, 하나의 웹하드 서비스로 분류한다.

전체적인 분류 알고리즘 (그림 2)와 같다. 시그니쳐를 등록한 XML 파일을 데이터화 한 후 학내 전체 트래픽 데이터에서 프로세스 코드와 시그니쳐를 이용해 분류해낸다.



(그림 2) 분류 시스템 구성도

5. 검증 방법

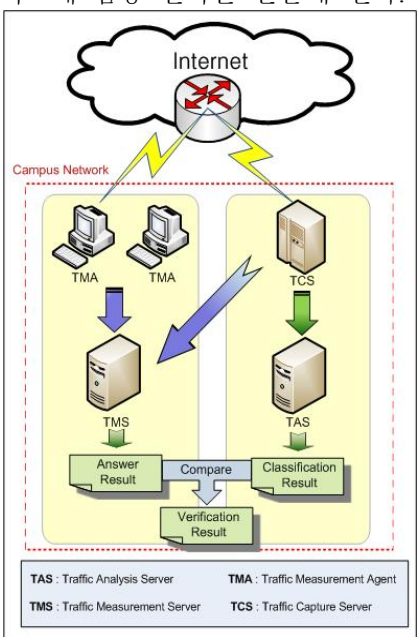
Flow 검증 시스템을 시그니쳐 생성 및 검증을 위해 사용하였다. 이 시스템은 signature 와 서비스, 프로세스 코드로 정형화된 XML 파일을 기반으로 Flow 들을 분류해내고, TMA[4] 기반으로 만들어진 정답지와 비교하여 트래픽 분류의 정확성을 웹페이지에 출력하게 된다. TMA 는 트래픽 분석과 검증을 위한 도구로써, 네트워크 내의 종단 호스트에 설치한다. TMA 는 매분 호스트의 소켓정보를 TMS(Traffic Measurement Server)로 전송시킨다. 모든 트래픽은 응용(application)에 의해 소켓에서 시작하고 종결됨으로 TMA 정보(Process name, IP Address, Port Number 등)와 Flow(인터넷과 연결된 백본 스위치에서 수집)를 비교하면 해당 트래픽을 발생 시킨 응용을 알아낼 수 있다.

정확도는 TP, TN, FP, FN 을 통하여 Precision 과 Recall 로 산출된다. 이 용어들에 대한 자세한 설명은 <표 5>를 통하여 알 수 있다.

<표 5>정확도 측정을 위해 사용된 용어

용어	설명
TP(X)	응용 X 의 트래픽을 올바르게 분류
TN(X)	응용 X 가 아닌 다른 응용(Y)의 트래픽을 올바르게 분류
FP(X)	다른 응용(Y)의 트래픽을 응용 X 의 트래픽으로 잘못 분류
FN(X)	응용 X 의 트래픽을 분류 못함
	응용 X 의 트래픽을 응용 Y 의 트래픽으로 잘못 분류
Precision	TP/(TP + FP)
Recall	TP/(TP + FN)

(그림 3)은 검증 시스템이다. TMA 에서 보내온 호스트들에 대한 정보를 분류 시스템에 의해서 분류된 플로우와 비교해 검증 결과를 산출해 낸다.



(그림 3) 검증 시스템 구성도

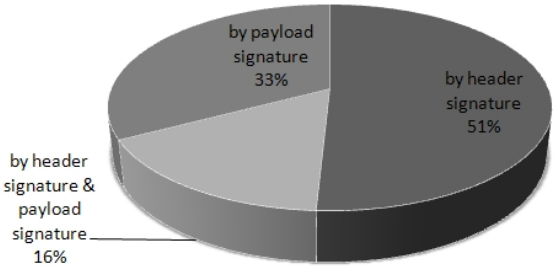
6. 실험 방법 및 실험 결과 분석

6.1 실험방법

분류의 검증 결과는 캠퍼스 내에서 인터넷으로 나가게 되는 혹은 들어오게 되는 모든 트래픽에 대하여 검증을 하였다. 현재 Flow 검증 시스템에서 정답지를 만들어내는 TMA 는 캠퍼스 내에 종단 호스트의 소켓 정보만 가지고 있으므로 검증의 결과로 캠퍼스내의 네트워크에 대한 분류의 정확성만이 알 수 있다.[2] 이 결과물은 현재 Flow Verification System 에서 지원되고 있는 2009 년 2 월 16 일의 하루의 보고서를 정리한 실험 결과이며 이를 토대로 분석하였다.

6.2 실험 결과 분석

웹하드 리스트는 header signature 와 payload signature 를 기준으로 61 개의 웹하드 서비스를 (그림 4)와 같이 그룹별 분류를 할 수 있다. (그림 4)에서 나타나듯이 상당수의 웹하드 트래픽은 header signature 만 가지고도 분류가 가능하였고, header signature 와 payload signature 가 같이 필요한 경우가 16%, payload signature 만 필요한 경우가 33%가 나왔다.



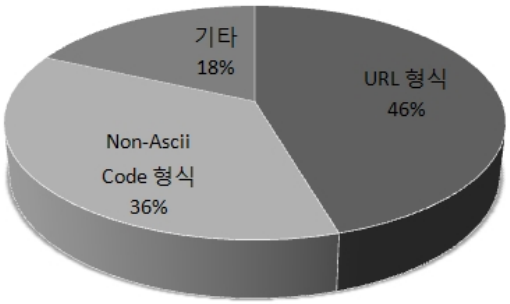
(그림 4) 시그니쳐 기준에 의한 분류

Header signature 에 의한 분류 시 remote ip 및 remote port 가 주를 이루었으며, payload signature 에 의한 분류 시 비슷한 형식의 payload signature 가 사용되었으며, <표 6>에서 사용빈도가 가장 높은 두 가지 형식을 나타내었고 (그림 5)에서 사용 빈도를 차트로 나타내었다.

<표 6> payload 데이터 형식과 그 예시

Payload 데이터 형식	예시
URL 형식	clubhard.co.kr
Non-Ascii Code 형식	01 00 00 00 CE FA 0B B0 A0 00 00 00 4D 4D 53

URL 형식은 대체적으로 그 웹하드 서비스의 URL 주소가 많이 나타났으며, Non-Ascii Code 형식은 특별한 패턴은 발견 되지 않았다. (그림 5)의 차트를 보면 URL 형식의 payload data 가 웹하드 서비스 분류에 있어서 1 순위로 쓰이는 것을 알 수 있다 지금까지 했던 분류들은 다음 <표 7>의 검증 결과와 <표 8>에 나와있는 검증을 통한 정확도 수치로 타당성을 입증할 수 있다.



(그림 5) Payload signature 형식에 의한 분류

<표 7>의 FP 수치와 다른 응용의 header signature 부분의 중복으로 일어났으며, FN 수치는 파일 다운로드 시 파일의 리스트를 받아오는 시점에서 대부분 나타나였다.

<표 7> 2 월 16 일 검증 결과

(단위 : flow)

Amount	TP	FP	FN
914442	889111	25331	1897

<표 8> 2 월 16 일 정확성 보고

Flow	Packet	Byte
97.22%	99.15%	99.63%

7. 결론 및 향후 과제

지금까지 signature 기반 방법으로 웹하드 트래픽을 분류하고 검증 하였으며, 검증 결과 95%이상의 높은 정확성으로 signature 기반 분류 방법으로 웹하드 서비스의 트래픽 분류가 상당 부분 가능하다는 것을 입증할 수 있었다.

향후 과제로는 빠른 속도로 많은 웹하드의 시그니쳐를 생성하는 시스템의 개발이 시급하며, 부하가 많이 걸리는 payload signature 대신 header signature 에 대한 연구를 통하여 header signature 의 동작 패턴 등의 특징을 이용한 분류 시스템 개발이 필요하다.

참고문헌

[1] 신재룡, 김경록, 곽윤식 “커스터마이징이 가능한 웹하드 디스크의 설계 및 구현”, 한국정보기술학회 하계학술대회 논문집, 2004. 8.

[2] 정혜원, 이준석, 서영호 “불법콘텐츠 추적 기술 연구동향”, 전자통신동향분석 제 20 권 제 4 호, 2005.8.

[3] 오영석, 박진완, 윤성호, 박준상, 김명섭 “시그니쳐 기반의 실시간 트래픽 분류 알고리즘의 성능 향상”, 통신정보 합동학술대회, 2009.

[4] SocketSniff, Website, http://www.nirsoft.net.

[5] 윤성호, 노현구, 김명섭, "TMA(Traffic Measurement Agent)를 이용한 인터넷 응용 트래픽 분류", 통신학회 하계종합학술발표회, 라마다플라자호텔, Jul. 2-4, 2008, pp.618.