HTTP 프로토콜을 이용한 웹 응용 정답지 트래픽 수집에 관한 연구

이태민, 구영훈, 이성호, 김명섭
고려대학교

{ltmsky, gyh0808, gaek5, tmskim}@korea.ac.kr

A Study on the Web Application Ground-Truth Collection using HTTP Protocol

Tae-Min Lee, Young-Hoon Goo, Sung-Ho Lee, Myung-Sup Kim
Korea Univ.

요 약

오늘날 휴대용 통신기기들의 발달과 네트워크 속도의 증가 및 다양한 웹 응용 서비스가 개발됨에 따라 웹 트래픽은 더욱 복잡해지며 다양해지고 있다. 인터넷에 기반을 둔 응용의 수와 트래픽이 급증하기 때문에 네트워크의 효율적인 관리를 위해 웹 트래픽 분류에 따른 트래픽 특성을 분석하는 것이 더욱 중요해지고 있다. 하지만 웹 응용 트래픽을 분류하는 많은 연구가 이루어지고 있음에도 불구하고 생성된 분류 결과에 대한 정확성을 증명하는 방법은 신뢰할 수 없다. 이는 특정 트래픽이 어떤 웹 응용에 의해서 발생했는지 정확하게 정의할 수 있는 Ground-Truth(정답지)의 기준이 없기 때문이다. 따라서 본 논문에서는 기준이 되어줄 정답지 트래픽을 HTTP 프로토콜의 Host 필드, Referer 필드, IP Address 를 이용하여 빠르고 정확하게 수집할 수 있는 방법을 제안한다. 실제로 웹 트래픽을 수집하고 제안된 방법론을 토대로 실험함으로써 그 타당성을 검증한다.

Ⅰ. 서 론

오늘날 휴대용 통신기기들의 발달과 네트워크의 고속화 및 다양한 소셜 네트워크의 발전으로 사람들은 기기를 통하여 언제 어디서든 인터넷에 접속해 원하는 정보를 요청하고, 공유할 수 있다[1]. 이로 인해 웹 트래픽은 더욱 빠르게 복잡해지며 다양해지고 있다. 이러한 환경 속에서 네트워크 관리를 위한 웹 트래픽의 정확한 분류에 따른 분석이 중요하기 때문에 현재 많은 분류 방법들이 연구되고 있다[2,3]. 그러나 웹 응용 트래픽을 분류하는 많은 연구가 이루어지고 있음에도 불구하고 생성된 분류 결과에 대한 정확성을 증명하는 방법은 신뢰할 수 없다. 이에, [1]본 논문은 HTTP 트래픽의 한 종류인 Request 패킷의 Host 필드, Referer 필드, Server IP Address 를 이용하여 분석 대상 망에서 수집하고자 하는 웹 응용들의 정답지 트래픽을 응용별로 빠르고 정확하게 수집하는 방법을 제안한다. 본 논문은 시중에 사용되고 있는 MS 에서 개발한 Network Monitor 프로그램을 사용한다. 하지만 Network Monitor 프로그램은 프로세스 별로 정답지 트래픽을 수집할 수 있을 뿐 응용 및 서비스까지 분류할 수 없기 때문에 본 논문에서는 응용과 서비스까지 분류할 수 있는 방법론을 제안한다.

본 논문은 1 장의 서론에 이어서 2 장에서는 시스템의 방법론을 간단한 예시를 통하여 기술한다. 3 장에서는 본 논문에서 제안하는 시스템을 토대로한 실험과 결과에 대해 기술하고, 마지막 4 장에서는 결론과 향후 연구에 대해 기술한다.

이 논문은 2015 년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구(No.2015R1D1A3A01018057) 및 2016 년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No.B0101-16-0300, 사이버 공격의 사전 사후 대응을 위한 사이버 블랙박스 및 통합 사이버보안 상황분석 기술 개발)

Ⅱ. 본론

본 장에서는 본 논문에서 제안한 시스템의 방법론에 대해서 기술하며, 시스템이 수행하는 과정을 간단한 예를 통하여 설명한다.

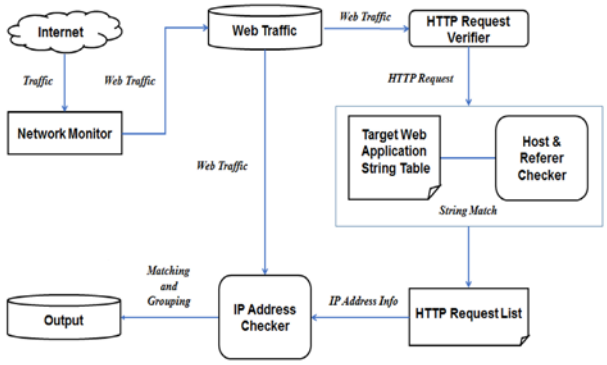

그림 1. 시스템 전체 흐름도

먼저, 본 논문에서는 웹 트래픽을 웹 브라우저에서 발생한 모든 트래픽을 의미한다. 그리고 본 논문에서 제안하는 정답지 트래픽은 3 계층 구조로 구분할 수 있다. 1 계층은 사용된 웹 브라우저의 종류, 2 계층은 웹 응용의 종류, 3 계층은 해당 웹 응용에서 제공하는 서비스를

의미한다. 전체 시스템의 흐름도는 그림 1 과 같다. Network Monitor 를 이용하여 트래픽 데이터를 수집한 후에 Network Monitor 에서 제공하는 기능 중 하나인 프로세스별 트래픽 분류를 이용하여 브라우저 별로 트래픽을 분류한다. 브라우저 별로 구분하는 이 단계에서 1 계층이 정해지게 된다. 분류된 트래픽을 저장한 다음 각 트래픽을 입력 데이터로 사용한다. 입력받은 트래픽에서 HTTP 의 Request 트래픽을 검출한다. Request 패킷을 검출하는 방법은 HTTP method 필드를 이용하는 것이다. HTTP 프로토콜 규정에서 정의된 method 는 <표 1>과 같이 총 8 가지이다.

GET	POST	HEAD	PUT
TRACE	CONNECT	DELETE	OPTIONS

<표 1> HTTP Method

Method 필드는 항상 HTTP Request 패킷 맨 앞 부분에 정의되어 있기 때문에, 실제 패킷의 맨 앞부분과 HTTP method 필드를 스트링 매치하여 일치하게 되면 해당 패킷이 HTTP Request 패킷이란 것을 알 수 있다[3]. HTTP Request 패킷을 이용하는 이유는 패킷 안의 Host 와 Referer 필드의 정보를 사용하기 위해서다. 두 필드에는 문자열 정보가 담겨있는데, 문자열 정보와 수집할 웹 응용 및 서비스에 대한 정보를 미리 테이블로 만들어 웹 응용 이름 및 서비스와 스트링 매치를 통해 해당 응용의 정답지 트래픽을 수집할 수 있다. HTTP Request 패킷을 검출한 다음엔 각 패킷의 Server IP Address, Host, Referer 정보를 중복 없이 리스트화한다. 아래의 <표 2>는 리스트화한 HTTP Request 패킷의 정보에 대한 예시를 보여주고 있다.

No.	Server IP	Host	Referer
1	104.118.6.108	Thumb.comic.naver.net	http://comic.naver.com...
2	211.244.82.171	Café.daum.net	http://cafe.daum.net...
3	121.125.76.70	www.afreecatv.com	http://www.afreecatv.com...
4	203.133.172.18	Map1.daumcdn.net	http://map.daum.net
5	23.32.241.152	Imgcomic.naver.net	http://comic.naver.com....

<표 2> Server IP, Host, Referer 필드 예시

리스트화한 정보에서 Host 또는 Referer 필드 안에 같은 응용의 이름 또는 같은 서비스에 관한 문자열이 나오면 해당 문자열을 가진 패킷끼리 차수가 낮은 계층에 우선순위를 두고 그룹을 짓는다. 마지막으로 그룹화된 패킷마다 각 패킷의 Server IP Address 정보를 이용하여 전체 트래픽과 IP Address 매치를 통해 리스트화 되어있는 그룹을 기준으로 분류한다. 이 단계에서 그룹화된 HTTP Request 패킷의 Server IP Address 정보를 사용하는 이유는 그룹화된 각 패킷과 전체 트래픽을 매칭 시켜 해당 IP Address 를 가진 패킷을 같은 그룹으로 포함시키기 위함이다.

Ⅲ. 실험 및 결과

이번 장에서는 본 논문에서 제안하는 수집 방법을 실험 결과를 통하여 그 타당성을 검증한다. 실험은 학내 망 안의 임의의 호스트에서 3 분간 발생한 트래픽을 수집하고 진행했다. 실험에 사용하기 위해 수집한 트래픽 정보는 <표 3>과 같다.

용량	전체 패킷 수
32,589 KB	48,786 개

<표 3> 실험에 사용한 트래픽 데이터

실험에 사용한 웹 응용은 Naver, Facebook, Afreeca TV, Daum 으로 네 가지 종류이다. <표 4>는 웹 트래픽과 HTTP 트래픽의 정보를 나타낸다. 트래픽을 수집한 후 Network Monitor 를 사용하여 1 차 구분한 Chrome 의 웹 트래픽은 28,344 개로 전체의 58%를 차지한다. 웹 트래픽에서 HTTP 패킷은 총 2,622 개지만, 2 계층 분류를 위해 첫 번째로 검출해낼 HTTP Request 패킷은 1,241 개다.

종류	Web traffic	HTTP	Request
수	28,344	2,622	1,241

<표 4> 웹 트래픽 및 HTTP 트래픽

1,241 개의 HTTP Request 패킷을 나열해보면 총 61 개의 IP Address 에서 발생했다. 61 개의 IP Address 에서 각 IP Address 마다 수집된 패킷들을 HTTP Request 의 IP Address 와 매칭시켜 그룹을 분류하여 수집된 패킷의 수는 23,580 개이다. 나머지 4,744 개 의 패킷은 웹 응용 정답지 트래픽으로 수집되지 못하였다. 수집하지 못한 응용의 트래픽은 대부분 암호화 되어있는 응용으로 많이 알려진 Facebook 과 CDN 서비스 트래픽으로 확인하였다. 결과적으로 1,241 개의 HTTP Request 패킷만으로 28,344 개의 웹 트래픽에서 약 83%인 23,580 개의 패킷을 응용별 정답지 트래픽으로 수집이 가능하였다.

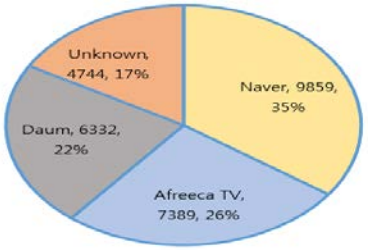

그림 2. 응용별 정답지 트래픽 수집 결과

Ⅳ. 결론

본 논문에서는 HTTP 프로토콜을 이용하여 웹 응용별로 정답지 트래픽을 수집하는 방법을 제안하였다. 또한 교내 망의 하나의 호스트에서 수집된 트래픽으로 실험을 진행하였고, 수집된 웹 트래픽 중 단지 4.3%에 해당하는 HTTP Request 패킷으로 전체 웹 트래픽에서 약 83% 이상을 웹 응용 정답지 트래픽으로 수집할 수 있는 것을 실험을 통해 증명하였다. 이렇게 수집한 정답지 트래픽은 특정 트래픽 혹은 패킷이 어떤 웹 응용에 의해서 발생하였는지에 대해 3 계층으로 이루어진 기준으로 정확한 정답지 기준을 제시할 수 있다.

향후 연구 계획으로는 본 논문에서 해결하지 못한 CDN 서비스 트래픽을 해당 응용의 정답지 트래픽으로 수집하는 방법과 제안한 방법론을 토대로 자동 웹 응용 정답지 트래픽 수집 방법을 연구할 계획이다.

참 고 문 헌

[1] 진창규, 김명섭, 최미정, “HTTP 트래픽의 서버별(사이트 측면) 분류 방법”, 2011 년 통신망운용관리 학술대회 (KNOM 2011), 포항공대, 한국, Apr. 21-22, 2011, pp. 18-21.

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

[3] 최지혁, 김명섭, “HTTP Host 를 이용한 웹 어플리케이션 인식에 관한 연구”, 정보처리학회 논문지 컴퓨터 및 통신시스템 Vol. 2 No8, Aug.2013, pp.327-334.