https://doi.org/10.7840/kics.2017.42.1.98

최신 네트워크 응용 분류를 위한 자동화 페이로드 시그니쳐 업데이트 시스템

심 규 석◆, 구 영 훈*, 이 성 호*, Baraka D. Sija*, 김 명 섭°

Automatic Payload Signature Update System for Classification of Recent Network Applications

Kyu-Seok Shim◆, Young-Hoon Goo*, Sung-Ho Lee*, Baraka D. Sija*, Myung-Sup Kim°

요 약

오늘날 네트워크 자원을 사용하는 응용이 증대되면서 네트워크 관리를 위한 트래픽 분석에서 현재 연구 단계의 한계가 드러나고 있다. 그런 한계를 해결하기 위한 다양한 연구가 진행되고 있는데 그 중 대표적인 연구인 시그니쳐 자동생성 연구는 응용 트래픽을 입력으로 트래픽의 공통된 패턴을 찾아 출력하는 과정이 자동화된 연구이다. 그러나 시그니쳐 자동생성 연구는 트래픽을 사용자가 수집해야 하는 반자동 시스템이기 때문에 트래픽 수집 단계에서 문제가 발생할 수 있고, 생성된 시그니쳐의 검증 과정이 포함되어있지 않기 때문에 시그니쳐의 정확도를 신뢰할 수 없는 한계가 있다. 본 논문에서는 시그니쳐 자동생성 시스템의 한계를 극복하기 위해 트래픽수집, 시그니쳐 생성, 시그니쳐 검증, 시그니쳐 관리까지 모든 과정이 자동으로 이루어지는 시스템을 제안한다. 제안하는 방법을 학내 망의 실제트래픽에 적용하여 추출한 시그니쳐는 분석률을 유지하며, 오탐률을 0으로 만드는 효과를 보였다.

Key Words : Automatic, Signature, Update, Generation, Management, Verification, Identifier

ABSTRACT

In these days, the increase of applications that highly use network resources has revealed the limitations of the current research phase from the traffic classification for network management. Various researches have been conducted to solutions for such limitations. The representative study is automatic finding of the common pattern of traffic. However, since the study of automatic signature generation is a semi-automatic system, users should collect the traffic. Therefore, these limitations cause problems in the traffic collection step leading to untrusted accuracy of the signature verification process because it does not contain any of the generated signature. In this paper, we propose an automated traffic collection, signature management, signature generation and signature verification process to overcome the limitations of the automatic signature update system. By applying the proposed method in the campus network, actual traffic signatures maintained the completeness with no false-positive.

※ 이 논문은 2015년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업과(No.2015R1D1A3A01018057) 2016년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 과학기술인문융합연구사업임(No.NRF-2016M3C1B6929228).
◆ First Author : Department of Computer and Information Science, Korea University, kusuk007@korea.ac.kr, 학생회원
° Corresponding Author : Department of Computer and Information Science, Korea University, tmskim@korea.ac.kr, 종신회원
* Department of Computer and Information Science, Korea University, gyh0808@korea.ac.kr, 학생회원
* Department of Computer and Information Science, Korea University, gaek5@korea.ac.kr, 학생회원
* Department of Computer and Information Science, Korea University, sijabarakajia25@korea.ac.kr, 학생회원
논문번호 : KICS2016-09-275, Received September 27, 2016; Revised December 16, 2016; Accepted January 13, 2017

Ⅰ. 서 론

네트워크 관리에 있어 가장 중요한 분야는 네트워크 모니터링이다. 네트워크 모니터링은 네트워크에서 특정 응용 및 서비스가 발생량을 알아내어, 그에 맞는 관리 정책을 수립하는 것을 의미한다. 네트워크모니터링을 위한 트래픽 분석은 사용자에게는 질높은 네트워크 서비스를 제공 받도록 하고, 제공자에게는 최소한의 네트워크 자원으로 최대한의 질 높은 서비스를 제공 할 수 있는 기반이 된다[1,2].

트래픽 분석에서 필수적으로 사용되는 것은 각 응용 별로 트래픽을 분류할 수 있는 시그니쳐 이다. 시그니쳐는 트래픽의 특징 별로 다양한 종류가 존재한다. 다양한 시그니쳐의 종류 중 본 논문에서는 정확도와 분석률이 가장 높은 페이로드 시그니쳐를 생성한다. 페이로드 시그니쳐는 트래픽의 데이터 부분인 페이로드 내에서 발생하는 고유하면서 연속된 문자열을 의미한다. 페이로드 시그니쳐를 이용하여 각 패킷의 페이로드를 검사하면서 매칭이 되었을 때, 해당 트래픽을 시그니쳐의 응용 별로 분류할 수 있다.

그러나 페이로드 시그니쳐를 생성하기 위해서는 많은 시간과 인력이 소비된다. 페이로드 시그니쳐를 생성하는 단계는 추출하고자 하는 응용의 트래픽을 수집한 뒤, 수집된 트래픽의 페이로드 내용을 직접 비교하면서 공통적으로 발생하는 문자열을 찾아 내야한다. 공통적으로 발생하는 문자열을 추출한 뒤 해당 응용에서만 발생하는지에 대한 검사를 진행한다. 따라서 추출하는 사람에 따라 시그니쳐의 품질이 달라질 수 있고, 이것은 시그니쳐의 객관성이 감소된다는 단점이 있다. 또한 추출 작업에 많은 시간이 소요되기 때문에 빈번하게 업데이트 되는 모든 응용에 대한 시그니쳐 업데이트는 사실상 불가능하다.

이러한 문제를 극복하기 위해 시그니쳐를 빠르고 정확하게 생성하는 시그니쳐 자동 생성 연구는 활발하게 진행되고 있다[3-8]. 시그니쳐 자동 생성 연구는 패킷의 페이로드 내용을 기반으로 공통 문자열을 자동으로 추출하여 시그니쳐 형태로 생성하는 방법이다. 그러나 시그니쳐를 만들기 위한 최소 조건은 추출하고자 하는 응용의 트래픽을 수집해야 하는데 이 단계는 사용자가 직접 트래픽 수집 도구를 이용하여 수집해야한다. 본 단계에서 트래픽을 잘 못 수집하게 되면 잘못된 시그니쳐가 추출 될 확률이 높다. 또한, 수집된 트래픽이 단기간의 트래픽이기 때문에 추출된 시그니쳐가 일회성의 성격을 가질 확률이 높다. 그리고 시그니쳐 검증 단계를 포함하지 않기 때문에 추출된 시그니쳐가 다른 응용을 분석할 확률이 높다. 마지막으로 항상 최신 시그니쳐를 유지할 수 없는 단점이 있다. 트래픽 패턴이 변화되더라도 인지하는 것은 여전히 사람이기 때문에 즉각적인 대응이 불가하다.

따라서 본 논문에서는 이러한 한계를 극복하기 위해 완전 자동화 시그니쳐 업데이트 시스템을 제안한다. 완전 자동화란 임의의 임력 없이 검증된 시그니쳐를 추출하고, 사용자는 시그니쳐를 지속적으로 업데이트 받을 수 있는 것을 의미한다. 제안하는 시스템은 트래픽 수집, 시그니쳐 생성, 생성된 시그니쳐 검증 그리고 시그니쳐 관리까지 모든 과정이 자동으로 이루어 지는 시스템이다. 본 시스템은 기존 시그니쳐 자동 생성 시스템의 단점을 해결할 수 있다. 반 자동 시스템으로 직접 트래픽을 수집해야하는 기존 시스템과 다르게 본 시스템은 TMA(Traffic Measurement Agent)에서 수집하는 프로세스 명칭과 5-tuple 및 시간 정보를 가지고 있는 Log 데이터와 트래픽 데이터를 비교하여 정답지 트래픽을 자동으로 생성하여 완전 자동화된 시스템으로 해결한다. 또한 일회성의 성격을 가질 시그니쳐는 지속적으로 시그니쳐 생성과정에서 선별되기 때문에 지속적인 시그니쳐 만을 추출할 수 있다. 오탐이 될 수 있는 시그니쳐는 다른 응용 트래픽과 비교하여 제거되기 때문에 낮은 오탐률을 가지는 시그니쳐를 추출할 수 있다. 마지막으로 본 시스템은 지속적으로 수행될 수 있기 때문에 트래픽 패턴이 변화되더라도 즉각적으로 변화된 시그니쳐를 추출할 수 있다.

본 논문은 1장 서론에 이어, 2장에서 시그니쳐 자동 생성 시스템에 대한 관련 연구에 대해 언급하고, 3장에서 제안하는 시스템을 제안한다. 4장에서는 제안한 시스템의 성능을 평가하고, 마지막 5장에서 결론 및 향후 연구에 대해 서술하고 논문을 마친다.

Ⅱ. 관련 연구

트래픽 분석을 위한 시그니쳐는 서론에서 언급했듯이 트래픽 특성에 따라 다양한 형태로 존재한다. 트래픽의 포트번호를 이용하여 분석하는 포트 기반 시그니쳐[9]와 트래픽의 크기, 위치, 시간 등 통계적인 정보를 이용한 통계 기반 시그니쳐[10], 패킷의 데이터 부분인 페이로드 정보를 이용한 페이로드 기반 시그니쳐[11]가 대표적이다.

본 연구에서는 페이로드 기반 시그니쳐를 다룬다. 페이로드 기반 시그니쳐는 패킷의 데이터 부분인 페이로드 내의 공통된 문자열을 의미한다. 가장 정확도

가 높은 시그니쳐이지만 시그니쳐 추출이 어렵고, 추출 과정에서 인적, 시간적 소비가 크다는 단점이 있다. 이러한 단점을 해결하기 위해 시그니쳐 자동 생성 방법이 연구되고 있다. 시그니쳐 자동 생성을 위해 다양한 알고리즘이 사용되고 있는데, 대표적으로 LCS (Longest Common String)알고리즘, Smith-Waterman 알고리즘, Autosig와 가장 최근 연구에서 순차 패턴 알고리즘의 한 종류인 AprioriAll 알고리즘을 이용한 시그니쳐 자동 생성 연구가 있다[3].

LCS 알고리즘을 응용 트래픽 시그니쳐 추출 목적에 맞게 변형한 대표적인 방법은 LASER (LCS-based Application Signature ExtRaction)이다. 본 방법은 두 개의 스트링을 비교하는 Matrix에서 Backtracking을 이용하여 연속된 공통 문자열을 찾는 방법이다. 따라서 두 개의 스트링을 계속 비교하여야 하기 때문에 추출 과정의 시간이 오래 걸리는 단점이 있다. 그림 1은 LCS 알고리즘을 이용하여 공통 문자열을 자동으로 찾아내는 LASER 방법이다. "payload"와 "payment" 두 개의 문자열이 있을 때, 다음과 같이 Matrix를 이용하여 Backtracking을 거쳐 "pay"라는 공통 문자열을 찾아낸다.

Smith-Waterman은 본래 DNA의 유사도를 판단하는 목적에 발표된 알고리즘이다[4]. 본 알고리즘을 사용한 응용 시그니쳐 자동 생성 방법도 발표되었는데 본 방법은 LCS와 매우 유사하다. 하지만 Backtracking 방법에서 LCS알고리즘은 연속된 공통 문자열을 찾을 수 있지만, Smith- Waterman 알고리즘은 연속된 공통 문자열의 집합을 찾을 수 있는 차이가 있다. 그러나 본 방법 또한 두 개의 스트링을 비교하는 것에 차이가 없기 때문에 추출 과정에 많은 시간이 소비된다.

가장 최근 연구인 AprioriAll 알고리즘[12,13]을 이용한 시그니쳐 자동 생성 방법은 위의 단점을 해결할 수 있는 방법이다. 위의 방법들은 특정 두 문자열을 비교하여 실제 트래픽에 적용하기 위해 트래픽의 순서를 정하거나, 그룹화 시키는 전처리 과정과 생성된 부분 문자열을 하나의 규칙으로 통합시키는 후처리 과정이 필요하다면, 본 방법은 모든 문자열을 후보로 길이1부터 증가시키며 시그니쳐가 될 수 있는 가능성이 높은 문자열만 취하기 때문에 추출 과정에 많은 시간이 소비되지 않고, 전처리 과정과 후처리 과정이 필요 없는 장점이 있다. 따라서 본 논문에서 시그니쳐 자동 생성 단계에서 AprioriAll 알고리즘을 사용한 시그니쳐 자동 생성 방법으로 시그니쳐를 추출한다.

그러나 위에서 언급한 기존 시그니쳐 자동 생성 시스템은 공통적인 한계가 존재한다. 첫 번째로 각 시스템 입력 데이터에 해당하는 각 응용의 트래픽을 사용자가 직접 수집해야하는 것이다. 두 번째는 트래픽을 사용자가 직접 수집하고, 시스템을 수행해서 추출되는 시그니쳐 이기 때문에 일회용 시그니쳐가 포함될 수 있다. 세 번째는 검증과정이 포함되어 있지 않아서 오탐 시그니쳐 추출 가능성이 있다. 마지막으로 최신의 시그니쳐를 항상 유지할 수 없다는 단점이 존재한다. 따라서 본 논문에서는 트래픽을 자동으로 수집하고, 시그니쳐 관리, 시그니쳐 생성, 그리고 시그니쳐 검증 단계를 모두 자동으로 수행하는 완전 자동화 시그니쳐 업데이트 시스템을 제안한다.

	0	1(p)	2(a)	3(y)	4(l)	5(o)	6(a)	7(d)
0	0	0	0	0	0	0	0	0
1(p)	0	D	←	←	←	←	←	←
2(a)	0	↑	D	←	←	←	←	←
3(y)	0	↑	↑	D	←	←	←	←
4(m)	0	↑	↑	↑	↑	↑	↑	↑
5(e)	0	↑	↑	↑	↑	↑	↑	↑
6(n)	0	↑	↑	↑	↑	↑	↑	↑
7(t)	0	↑	↑	↑	↑	↑	↑	↑

그림 1. LCS 알고리즘 공통 문자열 추출 방법
Fig. 1. The method of common sub-string extraction in LCS algorithm

Ⅲ. 완전 자동화 시그니쳐 업데이트 시스템

본 장에서는 완전 자동화 페이로드 시그니쳐 업데이트 시스템을 제안한다. 제안하는 시스템은 트래픽 수집, 시그니쳐 생성, 시그니쳐 검증, 그리고 시그니쳐 관리의 모든 과정이 자동으로 수행된다. 그림 2은 완전 자동화 페이로드 시그니쳐 업데이트 시스템의 수행 과정이다.

GT Traffic Generator: 자동 정답지 트래픽 수집

기존 연구가 가진 가장 큰 한계는 트래픽을 사용자가 직접 수집해야하는 것이다. 사용자는 트래픽 수집 도구를 이용하여 특정 응용의 순수한 정답지 트래픽만을 수집해야하는데 본 과정에서 다른 응용의 트래픽이 포함되거나 해당 응용의 트래픽을 다른 응용으로 판단하여 잘못된 트래픽이 수집될 수 있기 때문에 제안하는 시스템의 GT Traffic Generator에서는 트래픽을 자동으로 수집하고, 각 응용 별로 정답지 트래픽을 생성한다. 정답지 트래픽을 생성하기 위해 본 시스템에서는 TMA(Traffic Measurement Agent)[14]를 사

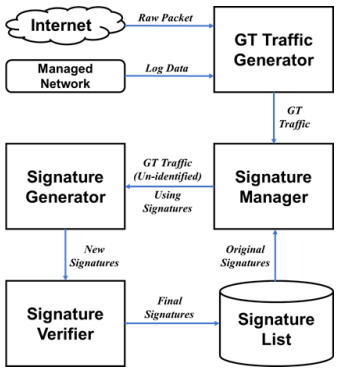


그림 2. 완전 자동화 페이로드 시그니쳐 업데이트 시스템 흐름도
Fig. 2. The flow of fully automatic payload signature update system

용한다. TMA는 각 호스트에서 실행되며 로그데이터를 남긴다. 로그 데이터에는 프로세스 이름, IP주소, 포트 번호, State, 프로토콜 등의 정보가 포함되어 있다. TMA는 각 호스트에서 시간대별로 TMS(Traffic Measurement Server)로 전송한다. TMS는 각 호스트에서 받아온 정보를 통합한다. 본 시스템에서는 TMS로부터 통합된 정보와 트래픽의 정보를 매칭한다. 같은 시간, 같은 5-tuple(srcIP/Port, dstIP/Port, Protocol)를 가진 정보를 매칭하여 트래픽을 프로세스 이름 별로 저장하며 응용 별 정답지 트래픽을 생성한다.

저장되는 트래픽은 시간단위로 저장되기 때문에 하루에 총 24개의 트래픽 파일이 수집된다. 기본 트래픽 파일은 분류가 안된 모든 트래픽을 의미한다. 그러나 TMA 로그 데이터와 트래픽 데이터를 매칭하여 각 프로세스 이름 별로 24개의 정답지 트래픽 데이터로 분류된다.

Signature Manager: 트래픽 분류 및 시그니쳐 관리

자동으로 수집된 정답지 트래픽을 이용하여 시그니쳐를 바로 생성할 수 있다. 하지만 같은 응용에 대해 지속적으로 동일한 시그니쳐가 추출될 수 있기 때문에 시스템에 불필요한 부하와 시간이 소비될 수 있다. 또한 기존 시그니쳐에서 다양한 원인으로 인해 사용되지 않는 시그니쳐에 대한 관리 방법이 필요하다. 본 시스템의 시그니쳐 관리단계에서는 기존 시그니쳐로 트래픽을 분석하여 분석되지 않는 트래픽을 분류하고, 사용되지 않는 시그니쳐를 삭제하며 시그니쳐를 관리한다.

최신 시그니쳐를 유지하기 위해서는 새로운 시그니쳐를 생성할 뿐만 아니라 사용되지 않는 시그니쳐를 삭제해야한다. 이러한 과정이 생략된다면 시그니쳐는 지속적으로 축적되고, 축적된 시그니쳐는 트래픽 분석에 있어 시스템 부하 및 과도한 시간 소비의 원인이 된다. 따라서 본 시스템에서는 사용되지 않는 시그니쳐를 삭제한다. 수식(1)은 시그니쳐의 구성을 나타낸다.

$$Signature=\{Header,Contents,Score\} \quad (1)$$

다음과 같이 시그니쳐는 응용 서버의 정보인 IP 주소, 포트 번호, 그리고 프로토콜로 이루어진 헤더정보와 트래픽의 고유한 패턴인 콘텐츠, 일회용 시그니쳐 삭제를 위한 Score값으로 구성된다. Score는 분석에 사용된 횟수를 의미하는데 시그니쳐 생성과 함께 초기값 1을 갖는다. 또한, 시그니쳐가 해당 트래픽의 분석에 사용되지 않으면 1이 감소되고, 분석에 사용되면 1이 증가한다. Score가 0이 되면 해당 시그니쳐는 삭제된다. 본 논문에서 Score을 사용하는 이유는 잘못된 트래픽 입력으로 인한 정상적인 시그니쳐가 삭제되는 것을 방지하고, 일회용 시그니쳐를 삭제하기 위함이다. Score는 분석에 지속적으로 사용되는 정상적인 시그니쳐는 누적적으로 Score가 증가되고, 비정상적인 일회용 시그니쳐의 경우는 Score가 감소되기 때문에 입력되는 트래픽이 잘못되더라도 정상적인 시그니쳐는 삭제되지 않을 수 있다. 그러나 Score가 매우 높으면 향후 트래픽 패턴이 변화되었을 때 삭제되지 않을 수 있기 때문에 Score은 최대 10으로 제한하여 향후 10번 연속으로 분석에 사용되지 않았을 경우 삭제된다.

Score값이 0값을 가지게 되면 해당 시그니쳐는 시그니쳐 리스트에서 삭제된다. 본 과정을 통해 삭제되는 시그니쳐는 크게 두 가지 타입으로 분류된다. 첫번째는 일회용 시그니쳐로 그 당시의 트래픽에서만 발생하는 시그니쳐이다. 대표적인 일회용 시그니쳐는 그림 5와 같이 날짜가 키워드로 포함된 시그니쳐이다. 이러한 일회용 시그니쳐는 그 날짜의 트래픽이 아니면 사용되지 않는 시그니쳐 이기 때문에 본 과정에서 삭제된다. 두번째 타입은 입력된 트래픽 데이터에서 우연적으로 발생하지 않은 시그니쳐이다. 예를 들어, AfreecaTV 응용에서 로그인에 대한 정상적인 시그니쳐가 있을 때, 다음 트래픽 분석에서 AfreecaTV 트래픽 중 로그인 트래픽이 발생하지 않다면 AfreecaTV의 로그인 시그니쳐는 일회용 시그니쳐로 분류할 수 있다. 이러한 오류를 방지하기 위해 Score을 이용한다. 따라서 본 과정을 통해 기존 시그니쳐로 분석되지 않은 트래픽과 기존 시그니쳐 중 일회용 시그니쳐가 삭제된 시그니쳐 Set이 출력된다.

Signature Generator: 시그니쳐 생성

시그니쳐 자동 생성은 시그니쳐를 추출하기 위해 순차 패턴 알고리즘 중 Apriorialll 알고리즘을 변형하여 사용한다[15]. 시그니쳐를 자동으로 생성하는 과정은 먼저 트래픽의 페이로드를 추출하여 시퀀스를 생성하고, 생성된 시퀀스들의 집합에서 길이1의 콘텐츠를 생성한다. 길이1 콘텐츠는 최소 지지도 검사를 통해, 길이2로 생성된 후보자 콘텐츠와 삭제될 콘텐츠로 구분된다. 더 이상 길이가 증가되지 않을 때까지 이러한 과정을 반복하여 공통 문자열을 추출한다.

생성되는 시그니쳐는 그림3와 같이 총 3가지 타입이 존재한다. 첫 번째 타입은 공통적으로 발생하는 연속된 문자열을 의미하는 콘텐츠 시그니쳐이다. 두 번째 타입은 동일한 패킷에서 발생하는 콘텐츠 시그니쳐의 조합을 의미하는 패킷 시그니쳐이다. 세 번째 타입은 동일한 플로우에서 발생하는 패킷 시그니쳐의 조합을 의미하는 플로우 시그니쳐이다.

콘텐츠 시그니쳐 추출 단계에서는 시퀀스 집합과 최소 지지도를 입력 받아 최소 지지도를 만족하는 콘텐츠를 추출한다. 본 과정에서 사용되는 알고리즘은 대용량 데이터베이스에서 순차 패턴을 찾는 AprioriAll 알고리즘을 콘텐츠 시그니쳐 추출 환경에 적용하였다. 알고리즘의 출력인 콘텐츠 SET은 여러 콘텐츠로 구성되어 있으며, 하나의 콘텐츠는 시퀀스 문자열의 연속된 부분 문자열을 의미한다.

길이를 1씩 증가시키면서 더 이상 새로운 콘텐츠가 추출되지 않을 때까지 콘텐츠 추출과 지지도 미만 콘텐츠 삭제 과정을 반복한다. 추출의 마지막 단계로써 추출된 모든 길이의 콘텐츠 포함관계를 확인하고, 만약 포함 관계에 있는 콘텐츠가 발견되면 해당 콘텐츠를 집합에서 삭제한다. 최종으로 생성된 콘텐츠 집합을 다음 단계인 패킷 시그니쳐 추출 단계로 전달한다. 다음 단계인 패킷 시그니쳐 추출 과정은 콘텐츠 시그니쳐 추출 과정과 크게 다르지 않다. 콘텐츠 시그니쳐 추출 과정에서 콘텐츠 시퀀스 구성할 때 트래픽 페이로드의 문자열로 시퀀스를 구성하였다면, 패킷 시그니쳐 추출은 트래픽 페이로드 문자열 대신 추출된 콘텐츠 시그니쳐가 사용된다. 그림 4은 패킷 시퀀스 추출 과정을 나타낸 것이다.

그림 4와 같이 패킷 시퀀스를 구성하고, 패킷 시그니쳐 추출 과정이 진행된다. 패킷 시그니쳐 추출 과정은 콘텐츠 시그니쳐 추출 과정과 동일하게 길이 1부터 길이 k까지 증가시키며 시그니쳐가 구성된다. 패킷 시그니쳐 추출 과정에서 길이 1인 패킷 시그니쳐 후보자는 추출된 콘텐츠 시그니쳐이다. 콘텐츠 시그니쳐를 위의 방법과 동일하게 조합하여 같은 패킷에서 발생하는 콘텐츠 시그니쳐 집합을 패킷 시그니쳐라고 한다.

다음 시퀀스를 이용해서 동일한 패킷 내에서 발생하는 콘텐츠 시그니쳐를 추출하고, 더 이상 추출되지 않을 때까지 본 과정을 반복한다. 패킷 시그니쳐가 추출되었을 때 포함관계 유무를 판단하여 포함 관계에 있는 콘텐츠 시그니쳐를 삭제하는 과정을 진행한다. 최종으로 생성된 패킷 시그니쳐 집합을 다음 단계인 플로우 시그니쳐 추출 단계로 전달한다.

플로우 시그니쳐 추출 과정은 패킷 시그니쳐 추출 과정에서 패킷 시퀀스를 구성하기 위해 콘텐츠 시그니쳐를 사용하였다면, 플로우 시그니쳐는 플로우 시퀀스를 구성하기 위해 패킷 시그니쳐를 사용한다. 따라서 플로우 시그니쳐는 같은 플로우에 존재하는 패킷 시그니쳐의 집합을 의미한다. 그림 5은 플로우 시퀀스 추출 과정이다. 그림 5와 같이 플로우 시퀀스를 구성하고, 플로우 시그니쳐 추출 과정을 진행한다.

플로우 시그니쳐는 전 단계에서 추출된 패킷 시그니쳐를 이용하여 같은 플로우 내에 존재하는 패킷 시

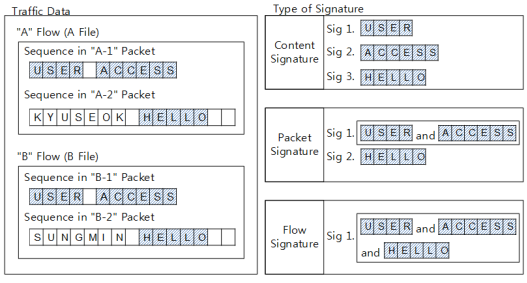


그림 3. 페이로드 시그니쳐 생성 타입
Fig. 3. The type of generated payload signature

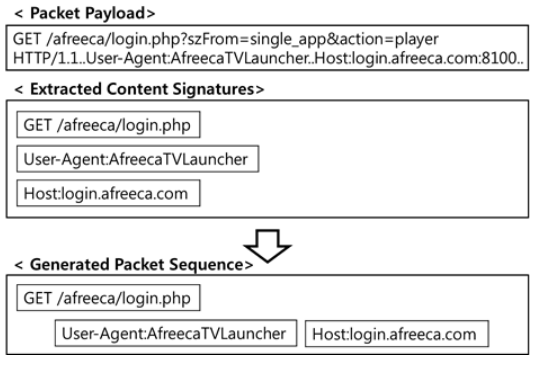


그림 4. 패킷 시퀀스 추출 과정
Fig. 4. The process of packet sequence extraction

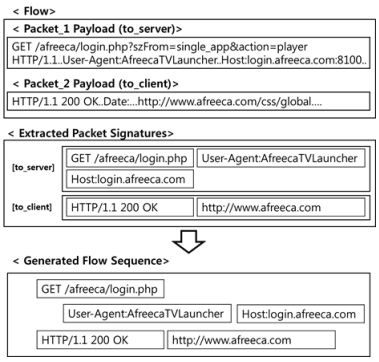


그림 5. 플로우 시퀀스 추출 과정
Fig. 5. The process of flow sequence extraction

그니쳐를 조합해서 생성된다. 플로우 시그니쳐의 길이를 점차 증가시키며 더 이상 길이가 증가되지 않을 때까지 반복한다. 플로우 시그니쳐를 생성하며, 포함관계에 있는 패킷 시그니쳐를 삭제한다. 최종적으로 플로우 시그니쳐 SET에는 포함관계에 포함되지 않은 콘텐츠 시그니쳐, 패킷 시그니쳐와 생성된 플로우 시그니쳐가 존재한다.

다음과 같이 공통으로 발생하는 연속적인 문자열을 의미하는 콘텐츠 시그니쳐 뿐만 아니라 패킷 시그니쳐, 플로우 시그니쳐를 생성하는 이유는 오탐률을 줄이기 위한 방법이다. 단순히 콘텐츠 시그니쳐만 추출된다면 사용자가 원하는 응용뿐 만 아니라 타 응용도 분류할 가능성이 높기 때문에 조건 검사에서 조건을 추가시켜 콘텐츠 시그니쳐 보다는 더 견고한 패킷 시그니쳐를 생성하고, 패킷 시그니쳐 보다 더 견고한 플로우 시그니쳐를 생성함으로써 오탐률이 적은 시그니쳐를 추출한다.

Signature Verifier: 시그니쳐 검증

제안하는 시스템은 위의 시그니쳐 생성과정에서 추출된 시그니쳐에 대해 검증 단계를 포함하고 있다. 오탐률이 적은 플로우 시그니쳐를 생성하지만, 플로우 시그니쳐도 오탐이 있을 수 있는 가능성이 있기 때문에 본 단계는 필수적인 단계이다. 다른 응용 트래픽을 분석하는 시그니쳐는 시그니쳐로써의 의미를 잃어버리기 때문에 삭제해주지만 아주 소량의 다른 응용 트래픽을 분석하고, 해당 응용 트래픽 분석에 큰 비중을 차지하고 있는 시그니쳐를 남기는 방법론을 제안한다.

시그니쳐 검증은 다른 응용을 분석하는 시그니쳐 즉, False-Positive가 있는 시그니쳐를 대상으로 한다. False-Positive가 없는 시그니쳐는 최종 시그니쳐에 포함된다. TP(True-Positive)는 A 응용 시그니쳐가 A응용 트래픽을 분석한 수치이다. FN(False-Negative)는 A응용 시그니쳐가 A응용 트래픽을 분석하지 못한 수치이다. FP(False-Positive)는 A응용 시그니쳐가 A 응용 트래픽을 제외한 나머지 트래픽을 분석한 수치이다. TN(True-Negative)는 A응용 시그니쳐가 A응용 트래픽을 제외한 나머지 트래픽을 분석하지 않은 수치이다.

$$\mathrm{Precision} = \frac{TP}{TP+FP} \tag{2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

F-measure는 Precision과 Recall을 이용하여 가중치를 주어 값을 측정하는 공식이다[16]. 본 논문에서는 F-measure를 사용하여 Precision에 가중치를 두기 위해 β값을 사용하는데 1을 기준으로 1보다 작으면 Precision에 가중치를 둔 수식이 되고, 1보다 크면 Recall값에 가중치를 둔 수식이 된다. 만약 동일한 가중치를 주어야한다면 β를 1로 고정하면 된다. 본 논문에서는 F-measure를 사용하여 Precision에 가중치를 두기 위해 β값을 0.1로 고정하여 사용한다. 수식(4)은 F-measure 표현식이다.

$$F-measure = \frac{(\beta^2+1)\times \mathrm{Precision}\times Recall}{\beta^2\times \mathrm{Precision} + Recall} \tag{4}$$

F-measure의 최대값은 1이고 최소값은 0이다. 1이 나오는 경우는 해당 시그니쳐가 해당 정답지 트래픽에 있는 모든 트래픽을 분석하고, 그 외 트래픽을 하나도 분석하지 못했을 때이고, 0이 나오는 경우는 TP=0가 되면 된다. 따라서 본 논문에서는 F-measure가 최소 0.95이상의 정확도를 가진 시그니쳐만을 최종 시그니쳐로 저장한다.

제안하는 방법을 통해 관리되는 네트워크에서 호스트가 사용하는 응용에 대한 시그니쳐는 지속적으로 업데이트가 되며, 사용되지 않은 시그니쳐는 삭제되고, 새로운 시그니쳐는 추출된다. 새로운 시그니쳐는 검증과정을 거치면서 정확도 높은 시그니쳐를 유지한다.

본 장에서는 트래픽 수집, 시그니쳐 관리, 시그니쳐 생성, 그리고 시그니쳐 검증까지의 과정을 모두 자동화한 완전 자동화 페이로드 시그니쳐 업데이트 시스템을 제안했다. 모든 과정이 자동화되어 있기 때문에 사용자 입장에서는 손쉽게 최신의 정확한 시그니쳐를 유지할 수 있다.

다음 그림 6과 같이 하나의 응용에 대해 자동 페이로드 시그니쳐 업데이트 시스템이 지속적으로 수행된다면, 정상적인 시그니쳐만 축적되는 기대효과를 나타낸 것이다. 시그니쳐 관리자를 통해 Score을 이용하여 일회용 시그니쳐는 삭제된다. 또한 시그니쳐 관리자에서 기존 시그니쳐로 분석하여 분석되지 않은 트래픽만 출력하기 때문에 추후 시그니쳐 생성자에서 같은 종류의 시그니쳐는 추출되지 않는다. 시그니쳐 생성자에서 추출되는 시그니쳐는 총 세가지 타입이 될 수 있다. 해당 응용 만을 분석할 수 있고, 일회성의 성격을 가지지 않은 정상적인 시그니쳐와 해당 트래픽에서만 적용되는 일회용 시그니쳐, 그리고 오탐이 가능한 시그니쳐이다. 그러나 본 단계에서 추출된 오탐이 가능한 시그니쳐는 다음 단계인 시그니쳐 검증 단계에서 삭제될 수 있다. 또한 일회용 시그니쳐는 다음 사이클에서 시그니쳐 관리 단계에서 삭제된다. 본 과정을 통해 사용자는 가장 최신의, 검증된 시그니쳐를 유지할 수 있다.

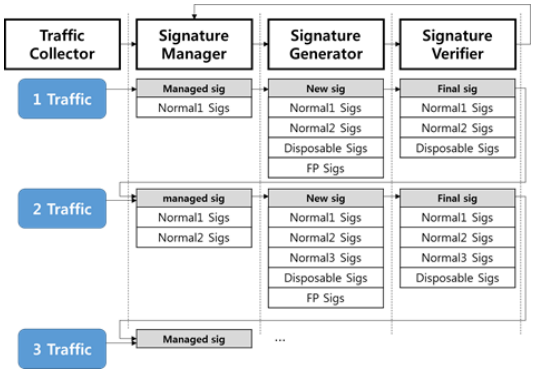


그림 6. 제안하는 방법의 수행 과정
Fig. 6. The process of proposed system

Ⅳ. 실험 및 성능 평가

본 장에서는 3장에서 제안한 방법인 완전 자동화 페이로드 시그니쳐 시스템의 성능을 평가한다. 자동으로 수집된 응용 트래픽 중 실험 및 평가를 위해 가장 많이 사용되는 4가지의 응용을 선정하여 실험을 진행하였다. 4가지의 응용은 동영상 및 방송 서비스를 제공하는 AfreecaTV[17], 소셜 네트워크 서비스를 제공하는 Facebook[18], 메신저 서비스를 제공하는 Kakaotalk[19], 그리고 파일 공유 및 전송 서비스를 제공하는 uTorrent[20]이다.

트래픽은 총 3가지 종류를 수집하였다. 초기 시그니쳐를 추출하기 위한 2015년 트래픽, 본 시스템 사이클을 위한 Update용 트래픽, 그리고 각 단계의 시그니쳐 분석률 및 오탐률을 평가하기 위한 평가용 트래픽으로 나누어서 수집되었다. 초기 시그니쳐 추출을 위해 2015년 트래픽을 사용한 이유는 1년 전 시그니쳐를 이용하여 올해 트래픽을 분석함으로써 최신 시그니쳐를 유지해야하는 중요성을 평가하기 위함이다. 업데이트용 트래픽과 평가용 트래픽은 2016년 5월 23일부터 26일까지 여러 호스트에서 발생한 트래픽을 자동 페이로드 시그니쳐 업데이트 시스템 중 정답지 트래픽 생성 단계에서 TMA의 로그데이터와 매칭하여 생성한 트래픽이다.

표 1은 각 응용 별 시스템 과정에 따른 시그니쳐의 분석률 및 오탐률을 나타낸 표이다. 분석률은 특정 응용의 전체 트래픽 중 해당 시그니쳐로 탐지된 트래픽의 비율을 의미하고, 오탐률은 특정 응용을 제외한 트래픽 중 해당 시그니쳐로 탐지된 트래픽의 비율을 의미한다. 표에서 Original Signature의 1_sig는 2015년 트래픽에서 추출한 시그니쳐이고, 해당 시그니쳐가 시그니쳐 관리 단계를 수행하면서 Using Signature의 1_sig가 된다. 이때 사용되는 트래픽은 첫번째 Update용 트래픽인 2016년 5월 23일 트래픽을 이용한다. Using Signature의 1_sig는 시그니쳐 관리 단계에서 출력된 분석되지 않은 트래픽을 이용해 시그니쳐 생성 단계에서 새로운 시그니쳐와 합쳐지면서 New Signature로 된다. 새롭게 추출된 시그니쳐와 사용 가능한 시그니쳐는 시그니쳐 검증 단계를 통해 오탐 가능성이 있는 시그니쳐를 제외한 나머지 시그니쳐로 Final Signature가 된다. 첫번째 Final Signature는 다음 사이클을 위해 Original Signature의 2_sig가 되서 다시 한번 이러한 과정을 수행한다.

전체적으로 Original Signature에서 시그니쳐 관리 단계를 지나 Using Signature로 변경될 때 시그니쳐의 개수는 급격히 줄어들었으나, 해당 응용을 분석하는 분석률은 큰 차이를 보이지 않는다.본 결과를 통해 시그니쳐 추출 과정에서 일회용 시그니쳐가 추출되는 사실을 확인했다.

다음 단계인 Using Signature에서 시그니쳐 생성 단계를 거쳐 New Signature로 변경될 때, 시그니쳐의 개수가 증가하고, 분석률도 증가하지만 오탐률도 증가하는 것을 확인하였다. 본 결과를 통해 시그니쳐 추출 과정에서 정상적인 시그니쳐도 추출되지만, 오탐이 가능한 시그니쳐도 추출되는 것을 확인하였다. 따라서 검증 과정이 필수적으로 필요하다는 것을 증명하였다. 다음 단계인 New Signature에서 시그니쳐 검증 단계

표 1. 완전 자동화 페이로드 시그니쳐 업데이트 시스템의 단계 별 시그니쳐 분석률 및 오탐률
Table 1. The completeness and false-positive rate of each signature of fully automatic payload signature update system

(Flows 단위: %)

		#Sig	Completeness				False-Positive			
			AfreecaTV	facebook	kakaotalk	torrent	AfreecaTV	facebook	kakaotalk	torrent
Original Sig	1	140	0.9	0.00	0.00	19	0.0	0.00	0.00	0.00
	2	56	30	25.3	33.7	36.7	0.0	0.16	0.00	0.00
	3	53	29.5	66.4	51.3	36.7	0.1	0.00	0.00	0.00
	4	44	37.8	78.3	88.3	36.7	0.0	0.05	0.08	0.00
Using Sig	1	2	0.9	0.00	0.00	19	0.0	0.00	0.00	0.00
	2	22	30	24.1	33.7	36.7	0.0	0.00	0.00	0.00
	3	28	27.5	65.4	51.3	36.7	0.0	0.00	0.00	0.00
	4	27	36.1	76.3	88.3	36.7	0.0	0.00	0.08	0.00
New Sig	1	59	31.1	95.6	50.1	36.7	0.1	31.2	0.99	0.00
	2	59	36.3	100	97.5	36.7	0.7	32.3	22.2	0.00
	3	49	42.8	85.8	88.3	36.7	1.4	0.39	0.01	0.00
	4	43	42.2	85.2	96.2	90.7	1.4	0.00	0.01	56.1
Final Sig	1	56	30	25.3	33.7	36.7	0.0	0.16	0.00	0.00
	2	53	29.5	66.4	51.3	36.7	0.1	0.00	0.00	0.00
	3	44	37.8	78.3	88.3	36.7	0.0	0.05	0.01	0.00
	4	40	36	85.2	96.2	41.1	0.0	0.00	0.01	0.00

를 통해 Final Signature로 변경되면서 분석률은 소량 감소하지만, 오탐률은 크게 감소하는 결과를 확인하였다. 따라서 검증 단계에서 오탐이 가능한 시그니쳐는 삭제되는 것을 확인하였다.

마지막 실험은 시그니쳐 검증 단계에서 F-measure 값을 사용하여 시그니쳐를 검증하고, 검증 결과에 따라 시그니쳐 삭제 여부를 판단하는데, F-measure값을 사용했을 때와 F-measure값을 사용하지 않았을 때의 차이를 확인한다. F-measure 값을 이용하는 이유는 해당 응용의 분석률에 영향이 크고, 오탐률이 존재하지만 미세한 시그니쳐를 삭제하지 않고 유지하기 위해서 이다. 해당 실험은 Kakaotalk 응용을 대상으로 실험을 진행하였다. 표 2, 3은 그 차이를 설명한다. New 시그니쳐 중 하나의 시그니쳐가 분석률에 큰 영향이 있지만, 오탐률이 있다. 본 시그니쳐가 F-measure

표 2. F-measure를 사용하지 않은 시그니쳐의 분석률 및 오탐률
Table 2. The completeness and false-positive rate of sigantures without F-measure

단위(%)

	#sig	Completeness			False-Positive		
		flow	pkt	byte	flow	pkt	Byte
New	28	96.2	76.1	72	0.08	0.01	0.00
Final	27	76.2	67.7	66.3	0.00	0.00	0.00

표 3. F-measure를 사용한 시그니쳐의 분석률 및 오탐률
Table 3. The completeness and false-positive rate of sigantures using F-measure

단위(%)

	#sig	Completeness			False-Positive		
		flow	pkt	byte	flow	pkt	Byte
New	28	96.2	76.1	72	0.08	0.01	0.00
Final	28	96.2	76.1	72	0.08	0.01	0.00

를 사용하지 않고, 오탐되어서 삭제된다면 오탐률은 0%로 만들 수 있지만 분석률은 급격히 감소되는 것을 확인하였다.

V. 결 론

본 논문에서는 트래픽 분류를 위한 최신의 정확한 시그니쳐를 유지하기 위해 트래픽 수집, 정답지 트래픽 분류, 시그니쳐 관리, 시그니쳐 생성, 그리고 시그니쳐 검증 단계의 모든 과정을 자동화하는 시스템을 제안하였다. 본 시스템은 기존 시그니쳐 자동 생성 시스템의 단점인 반자동 시스템 문제를 TMA&TMS를 이용하여 트래픽 수집 자동화하는 방법으로 해결하였고, 일회용 시그니쳐가 추출되는 한계는 시그니쳐 관리 단계에서 사용되는 시그니쳐만 선별하는 것으로 해결하였다. 그리고 높은 오탐률의 시그니쳐가 추출되

는 문제는 시그니쳐 검증과정을 통해 해결하였고, 최신의 시그니쳐 유지가 안되는 단점은 본 과정을 모두 자동화하여 지속적으로 시스템이 수행되어 항상 최신의 시그니쳐를 유지할 수 있는 방법으로 해결하였다. 본 논문에서 제안한 방법은 각 응용의 시그니쳐 중 정상적인 시그니쳐만 누적되고, 일회용 시그니쳐 및 오탐 시그니쳐는 삭제되며 시스템의 수행 횟수가 많아짐에 따라 응용의 분석률을 증가되고, 오탐률이 감소되는 것을 실험을 통해 확인하였다.

향후 연구로는 트래픽 수집부에서 TMA로그를 이용한 정답지 트래픽 분석은 각 호스트에 TMA가 설치되어야 정답지 트래픽 분류가 가능함으로 다른 방법을 통해 각 호스트에 TMA가 설치되어 있지 않더라도 정답지 트래픽 분류 가능한 방법을 적용할 예정이다. 또한 시그니쳐 생성 단계에서 현재 AprioriAll 알고리즘을 사용하여 시그니쳐를 생성하는데, 본 방법보다 더 최적화된 알고리즘을 적용하여 시스템의 속도를 향상시키고, 성능비교를 통해 해당 시스템의 성능 향상을 증명할 계획이다.

References

[1] M.-S. Kim, Y. J. Won, and J. W.-K. Hong, "Application-level traffic monitoring and an analysis on IP networks," *ETRI J.*, vol. 27, pp. 22-42, 2005.

[2] B. Park, Y. Won, J. Chung, M. S. Kim, and J. W. K. Hong, "Fine-grained traffic classification based on functional separation," *Int. J. Network Management*, vol. 23, pp. 350-381, Sept. 2013.

[3] B.-C. Park, Y. J. Won, M.-S. Kim, and J. W. Hong, "Towards automated application signature generation for traffic identification," *IEEE NOMS 2008,* pp. 160-167, 2008.

[4] X. Feng, X. Huang, X. Tian, and Y. Ma, "Automatic traffic signature extraction based on Smith-waterman algorithm for traffic classification," *IEEE Int. Conf. IC-BNMT*, pp. 154-158, 2010.

[5] H.-A. Kim and B. Karp, "Autograph: Toward automated, distributed worm signature detection," in *USENIX Security Symp.*, p. 19, San Diego, USA, Aug. 2004.

[6] Y. Wang, Y. Xiang, and S. Z. Yu, "An automatic application signature construction system for unknown traffic," *Concurrency and Computation-Practice & Experience*, vol. 22, pp. 1927-1944, Sept. 2010.

[7] Y. Choi, "An automated classifier generation system for application-level mobile traffic identification," 2011.

[8] P. Haffner, S. Sen, O. Spatscheck, and D. Wang, "ACAS: automated construction of application signatures," in *Proc. 2005 ACM SIGCOMM*, pp. 197-202, 2005.

[9] IANA port number list. Available: http://www.iana.org/assignments/service-names-port-numbers/service-names-port-numbers.xml

[10] N. F. Huang, G. Y. Jai, H. C. Chao, Y. J. Tzang, and H. Y. Chang, "Application traffic classification at the early stage by characterizing application rounds," *Inf. Sci.*, vol. 232, pp. 130-142, May 2013.

[11] T. Choi, C. Kim, S. Yoon, J. Park, B. Lee, H. Kim, et al., "Content-aware internet application traffic measurement and analysis," *IEEE/IFIP NOMS 2004*, pp. 511-524, 2004.

[12] R. Agrawal and R. Srikant, "Mining sequential patterns," in *Proc. 11th Int. Conf. Data Eng.*, pp. 3-14, 1995.

[13] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. 20th Int. Conf. VLDB*, pp. 487-499, 1994.

[14] S. H. Yoon, H. G. No, and M. S. Kim, "The classification of network application using the TMA," *KIPS Commun. 2008*, pp. 946-949, Daegu, Korea, May 2008.

[15] K. S. Shim, S. H. Yoon, S. K. Lee, S. M. Kim, W. S. Jung, and M. S. Kim, "Automatic generation of snort content rule for network traffic analysis," *J. KICS*, vol. 40, no. 04, pp. 666-677, Apr. 2015.

[16] D. M. W. Powers, "Evaluation: from precision, recall and f-measure to ROC, informedness, markedness and correlation," *J. Machine Learning Technol.*, vol. 2, no. 1, pp. 37-63, Dec. 2011.

[17] AfreecaTV, Available: http://www.afreecatv.com

[18] Facebook, Available: https://www.facebook.com

[19] Kakaotalk, Available: http://www.kakao.com/talk
[20] uTorrent, Available: http://www.utorrent.com

심 규 석 (Kyu-Seok Shim)

2014년 : 고려대학교 컴퓨터 정보학과 졸업
2016년 : 고려대학교 컴퓨터정보학과 석사과정 졸업
2016년~현재 : 고려대학교 컴퓨터정보학과 박사과정

<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

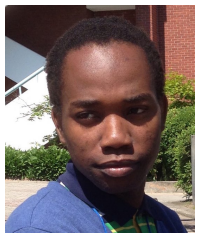
구 영 훈 (Young-Hoon Goo)

2016년 : 고려대학교 컴퓨터정보학과 학사
2016년~현재 : 고려대학교 컴퓨터정보학과 석사과정
<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

이 성 호 (Sung-Ho Lee)

2016년~현재 : 고려대학교 컴퓨터 정보학과 석사과정
<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석, 트래픽 분류

Baraka D. Sija



2016년~현재 : 고려대학교 컴퓨터 정보학과 석사과정
<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석, 트래픽 분류

김 명 섭 (Myung-Sup Kim)

1998년 : 포항공과대학교 전자계산학과 졸업
2000년 : 포항공과대학교 컴퓨터공학과 석사
2004년 : 포항공과대학교 컴퓨터공학과 박사
2006년 : Post-Doc. Dept. of ECE, Univ. of Toronto, Canada
2006년~현재 : 고려대학교 컴퓨터정보학과 부교수
<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석, 멀티미디어 네트워크