급변하는 네트워크 트래픽을 효율적으로 분류하기 위한 응용 시그니쳐 자동 생성 시스템에 대한 연구

심규석[1], 김종현[2], 김성민[3], 김명섭[0]
고려대학교 컴퓨터정보학과[1,3,0]
한국전자통신연구원[2]
{kusuk007[1], gogumiking[3], tmskim[0]}@korea.ac.kr
jhk@etri.re.kr[2]

A Study of Automatic Application Signature Generation System for Economically Classification Network Traffic of Dynamic Change.

Kyu-Seok Shim[1], Jong-Hyun Kim[2], Sung-Min Kim[3], Myung-Sup Kim[0]
Dept. of Computer and Information Science Korea University[1,3,0]
Electronics and Telecommunications Research Institute [2]

요 약

오늘날 다양한 응용 트래픽이 네트워크 환경에서 존재한다. 네트워크 관리자는 수없이 많은 종류의 네트워크 응용을 분류하기 위해 다양한 연구를 진행하고 있다. 그 중 하나인 시그니쳐를 이용한 네트워크 트래픽 분류 방법은 가장 많이 사용되고, 가장 신뢰도 있는 연구이다. 그러나, 시그니쳐를 이용하기 위해서는 정확하고 신속하게 시그니쳐를 추출해야만 하지만, 추출하기 위한 방법은 매우 시간적 소비와 인력적 소비가 매우 크다. 또한, 시그니쳐의 종류마다 장, 단점이 존재하기 때문에 확실한 시그니쳐를 찾기는 불가능하다. 따라서 본 논문에서는 정확하고 신속하게 네트워크 응용 시그니쳐를 추출하기 위해 헤더, 통계, 페이로드 정보를 이용한 복합 시그니쳐 자동 추출 시스템을 제안한다.

1. 서 론

오늘날 네트워크 환경에서는 통계적 추이가 불가할 정도로 많은 종류와 많은 양의 응용 트래픽이 발생하고 있다. 이러한 네트워크 응용 트래픽을 관리하기 위해 네트워크 관리자 및 연구원들은 다양한 방면으로 연구를 진행하고 있다[1]. 그러한 연구 중 하나인 시그니쳐를 이용한 네트워크 응용 트래픽 분류 방법은 가장 정확하고 대표적인 연구이다.

시그니쳐는 각 응용에서 발생하는 트래픽의 특성을 의미하기 때문에 다양한 종류의 시그니쳐가 존재한다. 먼저, 포트 기반 시그니쳐는 IANA에서 지정한 포트 정보를 이용한 트래픽 분석방법이다[2]. 해당 방법은 많은 응용들이 방화벽 및 IPS 장비를 통과하기 위해 포트 번호를 임의로 설정하여 트래픽을 발생시키기 때문에 현재는 많이 사용하지 않는 방법이다. 두번째는 헤더 정보를 이용한 시그니쳐이다[3]. 헤더 정보는 서버의 IP 주소와 port 번호를 의미한다. 해당 방법은 매우 빠르게 응용을 분류할 수 있는 장점이 있지만, 분류 범위가 넓고 정확하지 않다는 단점이 있다. 세번째는 페이로드 기반 시그니쳐이다[4]. 페이로드는 트래픽에서 데이터부를 의미한다. 각 응용 트래픽의 데이터부에서 공통적으로 발생하는 유일한 문자열이 페이로드 시그니쳐이다. 해당 방법은 매우 정확하다는 장점이 있지만, 시그니쳐 추출 과정이 매우 까다롭고, 시간/인력 소비가 매우 크면서 암호화된 트래픽을 분류할 수 없는 단점이 있다. 네번째는 통계 기반 시그니쳐이다[5]. 각 응용 트래픽의 패킷 사이즈, 시간등의 통계적 정보를 이용하여 응용 트래픽을 분류하는 방법이다. 해당 방법은 페이로드 시그니쳐로 분류하지 못하는 암호화된 트래픽을 분류할 수 있는 장점이 있지만, 통계 기반 시그니쳐를 추출할 수 있는 응용이 한정적이고, 정확도 면에서 페이로드 시그니쳐와 비교하여 낮다는 단점이 있다.

현재까지 모든 연구는 각 시그니쳐 종류의 단점을 해결하기 위한 연구는 다양하게 진행되어 왔다. 대표적인 연구는 생성과정이 매우 복잡하고, 까다로운 페이로드 시그니쳐를 자동으로 생성하는 연구이다. 하지만, 하나의 시그니쳐 종류로 모든 단점을 해결할 수 없다. 페이로드 시그니쳐를 자동으로 생성하는 연구는 생성 과정의 단순화 할 수 있지만, 암호화된 트래픽을 분류할 수 없는 단점은 여전히 존재한다.

따라서, 본 논문에서는 헤더, 통계, 페이로드 기반

이 논문은 2015년도 정부(교육부)의 재원으로 한국연구재단의 지원(No.2015R1D1A3A01018057)과 2016년도 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)을 받아 수행된 기초연구사업임.

시그니쳐의 장점을 최대로 활용하기 위해 복합 시그니쳐 자동으로 생성하는 시스템을 제안한다. 본 시스템은 시그니쳐 중 가장 가볍고, 분류시간이 가장 짧은 헤더 시그니쳐를 생성하고, 헤더 시그니쳐로 분류하지 못한 트래픽을 위해 페이로드 시그니쳐를 두번째로 생성한다. 마지막으로 페이로드 시그니쳐로 분류하지 못한 암호화된 트래픽을 분류하기 위해 통계 시그니쳐를 생성한다.

본 논문은 1장 서론에 이어, 2장에서는 복합 시그니쳐의 생성 과정을 설명한다. 마지막으로 3장에서 결론 및 향후연구에 대해 언급한다.

2. 본 론

본 장에서는 앞서 언급한 세 종류의 시그니쳐인 헤더, 페이로드, 통계 기반 시그니쳐의 단점을 최소화하고, 장점을 최대화할 수 있는 복합 시그니쳐의 자동 생성 시스템에 대해 언급한다. 먼저, 복합 시그니쳐는 헤더, 페이로드, 통계 기반 시그니쳐를 의미한다. 헤더 정보 기반 시그니쳐는 트래픽을 빠르게 분류할 수 있는 큰 장점이 있다. 그러나 신뢰도 있는 분류가 아니고, 분류된 트래픽의 범위가 매우 크기 때문에 분류된 내용에서 다시 페이로드 기반 시그니쳐를 이용하여, 신뢰도 있는 분류 과정이 진행된다. 그러나, 암호화된 트래픽의 경우 페이로드 시그니쳐로 분류할 수 없기 때문에 통계 기반 시그니쳐를 이용하여 트래픽을 분류한다. 그림 1은 복합 시그니쳐가 트래픽을 분류하는 과정을 표현한 것이다.

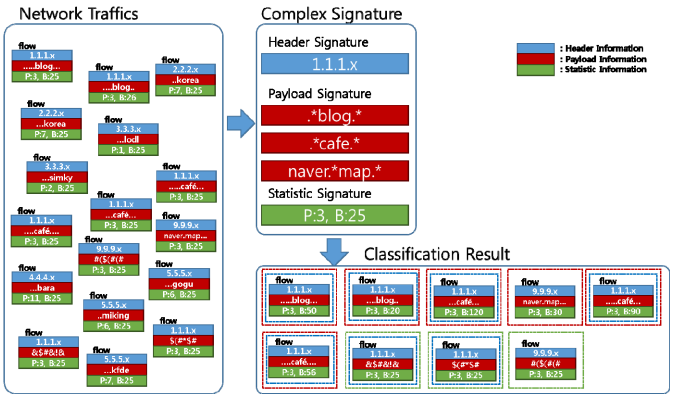

그림1. 복합 시그니쳐를 이용한 트래픽 분류

그림 1과 같이 헤더, 페이로드, 통계 정보를 포함한 복합 시그니쳐는 헤더 시그니쳐로 트래픽을 빠르게 분류하고, 페이로드 시그니쳐로 분류된 트래픽은 더 상세한 분류 과정을 가지고, 헤더 시그니쳐로 분류되지 않았던 트래픽을 분류할 수 있다. 또한 페이로드 시그니쳐로 분류되지 않은 암호화된 트래픽 또한 통계 시그니쳐를 이용하여 분류할 수 있다. 따라서 특정 응용이 발생하는 모든 형태의 트래픽을 분류할 수 있다. 해당 그림은 모든 경우를 보여주기 위한 하나의 예이다. 실제 트래픽에 적용하였을 때, 암호화된 트래픽을 발생하는 응용에 대해서는 통계 시그니쳐가 주로 사용되고, 헤더 정보로는 분류하지 못하는 트래픽이 많이 발생한다. 본 복합 시그니쳐는 어떤 응용 트래픽에도 적용할 수 있는 시그니쳐이다.

시그니쳐의 장점이 극대화 되고, 단점을 해소할 수 있는 복합 시그니쳐를 자동으로 추출하기 위해 본 논문에서는 복합 시그니쳐 자동 생성 시스템을 제안한다. 본 시스템은 특정 응용의 트래픽을 입력으로, 트래픽의 다양한 정보를 활용하여 헤더 정보를 이용한 헤더 시그니쳐, 페이로드를 이용한 페이로드 시그니쳐, 그리고 통계 정보를 이용한 통계 정보 시그니쳐를 추출한다. 이러한 방법은 네트워크 트래픽의 한 부분만을 초점으로 시그니쳐를 생성하는 것이 아닌, 모든 정보를 활용하여 시그니쳐를 추출하는 만큼 다양한 형태로 발생하는 트래픽을 대부분 높은 분석율과 정확도로 분류할 수 있다는 장점이 있다. 그림 2는 복합 시그니쳐 자동 생성 시스템의 시그니쳐 추출 과정을 표현한 그림이다.

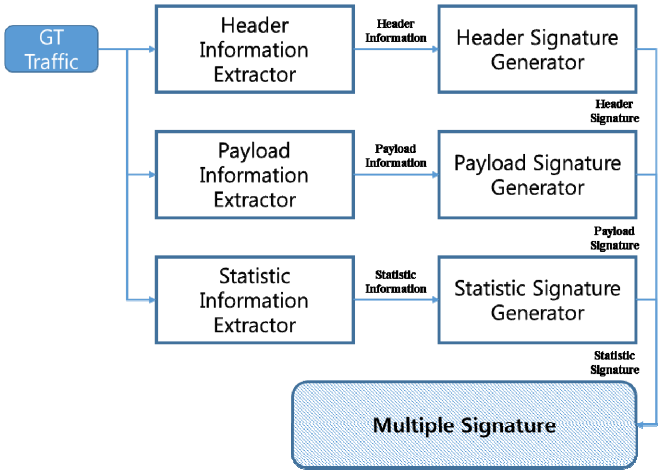

그림2. 복합 시그니쳐 자동 생성 시스템

그림 2와 같이 복합 시그니쳐 자동 생성 시스템은 총 6개의 단계를 거친다. 먼저 순수한 특정 응용의 트래픽만 수집되어 있는 정답지 트래픽을 이용하여, 해당 트래픽의 헤더 정보를 추출한다. 헤더 정보는 3-tuple(서버 측의 IP와 포트번호 L4 프로토콜)을 추출한다. 추출된 3-tuple 데이터를 이용하여 헤더 시그니쳐를 생성한다.

페이로드 시그니쳐는 페이로드 정보 추출기를 통해 추출된 페이로드 정보를 이용하여 생성된다. 페이로드 정보 추출기에서는 트래픽의 데이터부인 페이로드를 추출하고, 페이로드 시그니쳐 생성부에서는 추출된 페이로드를 이용하여 공통된 문자열을 찾아 페이로드 시그니쳐를 생성한다. 페이로드 시그니쳐는 플로우 단위로 생성되게 되는데 이때, 순차 패턴 알고리즘의

종류 중 하나인 AprioriAll 알고리즘을 사용하여 공통 문자열을 찾고, 패킷 단위에서 발생하는 문자열 집합을 찾고, 플로우 단위에서 발생하는 패킷 페이로드 시그니쳐를 찾아내어 플로우 단위 페이로드 시그니쳐를 추출한다.

통계 기반 시그니쳐를 추출하기 위해서는 통계 정보 추출기에서 통계 정보를 추출한다. 통계 정보에는 패킷의 방향, 크기, 순서 등이 포함된다. 이러한 통계 정보를 이용하여 통계 시그니쳐 생성부에서는 플로우의 통계 정보를 각각 벡터로 표현하여, 그룹핑한다. 그룹핑 된 플로우들의 유사도를 측정하여, 최종적으로 통계 기반 시그니쳐를 추출한다.

3. 결 론

본 논문에서는 다양한 형태로 발생하는 네트워크 응용 트래픽을 분류하기 위해 복합 시그니쳐와 복합 시그니쳐를 자동으로 생성하는 시스템을 제안했다. 복합 시그니쳐는 헤더, 페이로드, 통계 정보를 모두 포함하고 있어서, 정확하게 트래픽을 분류할 수 있을 뿐만 아니라 기존에 페이로드 시그니쳐로 분류가 불가능했던 암호화된 트래픽을 분류할 수 있다. 또한, 복합 시그니쳐를 자동으로 생성하는 시스템을 제안함으로써 시그니쳐 생성 또한 간편하게 할 수 있다.

향후 본 시스템의 성능을 평가하고, 생성된 시그니쳐를 간편화하는 연구를 할 예정이다.

4. 참고문헌

[1] Y. Wang, Y. Xiang, W. L. Zhou, and S. Z. Yu, "Generating regular expression signatures for network traffic classification in trusted network management," Journal of Network and Computer Applications, vol. 35, pp. 992-1000, May 2012.

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

[3] Sung-Ho Yoon, Myung-Sup Kim, "An Efficient Method to Maintain the Header Signature for Internet Traffic Identification," Proc. of the Asia-Pacific Network Operations and Management Symposium (APNOMS) 2013, Hiroshima, Japan, Sep. 25-27, 2013.

[4] Jun-Sang Park, Sung-Ho Yoon, and Myung-Sup Kim, "Software Architecture for a Lightweight Payload Signature-based Traffic Classification System," Proc. of the Traffic Monitoring and Analysis (TMA) Workshop 2011, LNCS6613, Vienna, Austria, Apr. 27, 2011, pp. 136-149.

[5] Hyun-Min An, Su-Kang Lee, Jae-Hyun Ham, and Myung-Sup Kim, "Traffic Identification based on Applications using Statistical Signature free from Abnormal TCP Behavior," JOURNAL OF INFORMATION SCIENCE AND ENGINEERING, Vol.31 No.5, Sep. 2015, pp. 1669-1692.