# 어플리케이션의 네트워크 행동 기반 시그니처 자동 생성 및 수집

정재윤, 원영준, 홍원기
포항공과대학교 컴퓨터공학과
{dejavu94, yjwon, jwkhong}@postech.ac.kr

# An Automated Behavior Signature Generation System for Application Traffic Identification

Jae Yoon Chung, Young J. Won, and James Won-Ki Hong
Dept. of Computer Science and Engineering, POSTECH

## 요 약

본 논문은 트래픽에서 응용프로그램의 네트워크 행동 시그니처 자동 생성과 생성된 행동 시그니처의 수집에 관한 것으로, 각 응용프로그램의 네트워크 사용의 특징을 분석하는 방법과 찾아낸 행동 시그니처를 요약하여 수집 및 정리하는 방법을 제시한다. 최종 사용자의 호스트에 위치하여 각 응용프로그램의 네트워크 행동 시그니처를 추출하는 에이전트 프로그램과 추출한 시그니처를 통합 관리하는 서버를 개발하였다. 제시하는 방법은 응용프로그램이 비정상적인 네트워크 행동을 보일 경우 이를 탐지 해 낼 수 있으며, 악의적인 행동을 하는 응용프로그램을 탐지 할 수 있다. 또한 여러 에이전트 프로그램에서 수집된 행동 시그니처를 서버에 수집하고 공개함으로써 시그니처의 개체 수 및 정확도를 높이고 그 접근성을 높일 수 있다.

## I. 서론

현재 우리가 사용하는 거의 모든 응용프로그램은 인터넷 등의 네트워크 자원을 사용한다. 수많은 응용프로그램의 트래픽이 네트워크를 돌아다니면서 그들이 의도하던 의도하지 않던 종종 여러 가지 문제를 발생시킨다. 네트워크 관리자는 점차 각각의 응용프로그램이 네트워크에 어떠한 영향을 끼치고 있으며 그 특성은 무엇인지에 대한 정보가 필요하게 되었다.

가장 일반적으로 트래픽을 분류하는 방법은 포트(port)기반의 분류방법[1]이다. 하지만 이 방법은 방화벽이나 보안 시스템을 피하기 위해 동적 포트를 사용하거나 well-known 포트를 자신만의 목적으로 사용하는 여러 응용프로그램이 등장하면서 트래픽 분류의 정확도가 약 70%밖에 되지 않는다[2].

이러한 단점을 극복하기 위하여 패킷의 페이로드(payload)에서 추출한 시그니처(signature)를 기반으로 한 연구가 진행되었다[3, 4]. 시그니처는 어떤 트래픽 플로우(flow)가 어떤 인터넷 응용프로그램으로부터 발생하였는지를 알아내기 위한 식별자(identifier)로서, 트래픽 플로우를 구성하는 패킷의 페이로드로부터 추출되는 정보이다. 하지만 이러한 종래 기술에 의하면 패킷의 페이로드를 모두 조사해 봐야 하는 부담이 있고, 이는 네트워크 최종 사용자가 원하지 않는 문제점이 있다. 또 페이로드의 시그니처 기반의 방법은 미리 등록된 시그니처가 없는 트래픽은 분류할 수 없다는 단점이 있다.

기존의 행동 기반 트래픽 분류 방법[5, 6]은 패킷의 해더 정보를 이용하여 오버레이 네트워크의 형태를 살펴보거나 각각의 응용프로그램 사이의 네트워크 연결 상태를 살펴보는 방법이다. 대부분이 비정상 트래픽의 탐지에 사용되고 있다. 행동 기반 트래픽 분류 방법의 단점으로 응용프로그램의 행동을 묘사하기 힘들다는 점과 통계 기반의 접근법에 따른 많은 training 정보의 필요가 단점으로 지적되고 있다.

우리는 이번 연구를 통해 시그니처를 자동 생성하는 에이전트와 이를 수집하는 서버, 시그니처를 요약 표현하는 XML 문서를 정의하였고 이 세가지를 모두 포함하는 통합 시스템을 제시한다.

## II. 본론

앞서 말한 바와 같이 행동 기반의 시그니처 자동 생성 및 수집 시스템은 최종 사용자(end user)의 위치에서 응용프로그램의 시그니처를 추출 분석하는 에이전트 프로그램과 이들 에이전트 프로그램으로부터 요약된 응용프로그램의 행동 시그니처를 전송 받아 통합 관리하는 시그니처 수집 서버로 구성된다. 아울러 이 연구는 에이전트와 서버 사이에 오가는 행동 시그니처 정보의 요약 표현 방법도 정의한다.

### 1. 시그니처 수집 에이전트

시그니처 수집 에이전트는 개략적으로 그림 1 처럼 구성된다. 최종 사용자가 주고 받는 패킷을 볼 수 있는 패킷 수신 장치 부와 패킷의 해더 및 페이로드를 나눠주는 파싱부가 있다. 패킷의 해더 정보와 페이로드 정보는 등록된 행동 시그니처 패턴들과 비교되어 행동 판단부로 넘겨진다. 트래픽 행동 판단부는 등록된 패턴들을 바탕으로 응용프로그램의 행동 시그니처를 찾아내는 가장 핵심적인 부분이다. 생성된 시그니처는 전송에 알맞은 형태로 요약되며 이를 시그니처 스냅샷이라 부른다. 에이전트의 마지막 구성 요소는 이 스냅샷을 서버로 전송시켜주는 부분이다.

"본 연구는 지식경제부 및 정보통신연구진흥원의 IT 핵심기술개발사업의 일환으로 수행하였음" [2008-F-016-02, 초정밀 측정 및 분석 기술 연구]

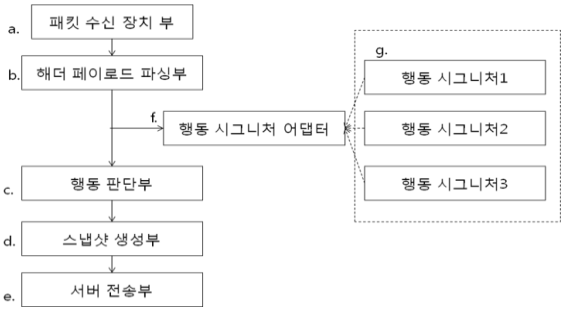


**그림 1. 시그니처 생성 에이전트 구성**

에이전트는 그림 2 의 흐름을 따라 수행된다. 버퍼에 쌓인 패킷을 읽어 필요한 정보를 얻은 후 pluggable 하게 추가/제거 할 수 있는 행동 시그니처 plug-in 에 해당하는 행동 시그니처를 추출한다. 이때 프로세스 별 플로우의 구분이 우선 수행되어야만 한다. 행동 시그니처 plug-in 에 다라 추출된 행동 시그니처는 XML 문서형태로 요약 및 표현되어 시그니처 수집 서버로 전송된다. 이 XML 문서는 Component 라고 불리는 sub-tree 를 갖고 있는데 각각의 Component 는 각각의 행동 시그니처 plug-in 이 표현하고자 하는 정보를 최대한 유연하게 표현하도록 정의되었다.

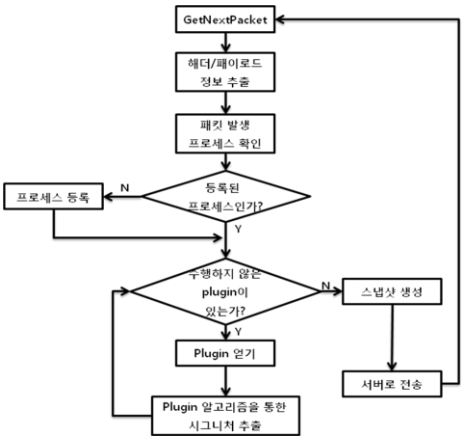


**그림 2. 에이전트 흐름도**

2. 시그니처 수집 서버

그림 3 은 시그니처 수집 서버의 전체적인 구성을 그리고 있다. 시그니처 수집 서버는 에이전트로부터 요약된 시그니처 정보인 스냅샷을 수신하는 부분과 스냅샷의 인증 및 기존 시그니처 스냅샷과 비교하는 스냅샷 수집부, 그리고 시그니처 추출 알고리즘을 이용하여 응용프로그램별 통합 시그니처를 추출하는 시그니처 추출부로 구성된다. 시그니처 추출 알고리즘으로 기계학습 기반의 행동 패턴 분석 방법이나 통계 기반의 패턴 인식 알고리즘이 사용 될 수 있다.

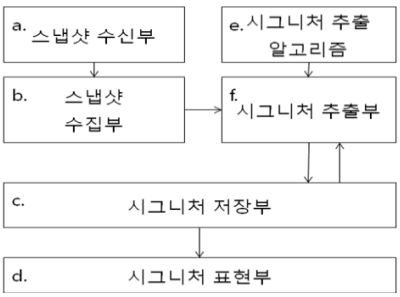


**그림 3. 시그니처 수집 서버 구성**

서버에 수집된 시그니처는 서버와 에이전트 사이에서 주고 받는 시그니처 스냅샷 형태로 요약되어 외부에 공개된다.

3. 시그니처 스냅샷

시그니처 스냅샷은 XML 문서 형태로 표현된다. 시그니처 정보를 XML 문서 형태로 보다 유연하게 표현하기 위해서 각각의 시그니처 패턴을 Component 라는 단위로 표현하였다. 각 Component 는 플로우 정보, 플로우당 패킷 수, 플로우당 바이트 등의 정보를 갖고 있으며 한 Component 안에 여러 플로우 정보가 있을 경우 이를 분석한 추가적인 정보(e.g. linear scanning, random scanning, and etc)들도 포함한다. 예를 들어 에이전트에 worm 을 탐지하기 위한 plug-in 이 있는 경우, 에이전트는 worm traffic 을 분석하여 worm 프로세스가 주변 네트워크 자원을 찾기 위한 스캐닝 정보를 추출하여 Component 에 넣어준다.

## III. 결론

이번 연구에서 우리는 행동기반 시그니처를 자동으로 추출하고 이를 수집하는 시스템을 제시하였다. 에이전트 프로그램은 호스트에 설치되어 최대한 효율적으로 시그니처를 추출하여 서버로 시그니처 스냅샷을 전송하고 서버는 이를 수집하여 보다 정확한 네트워크 행동 기반 시그니처를 제시한다. 에이전트와 서버 사이의 전송 포맷은 XML document 형태로 하였으며 다양한 행동 기반 시그니처를 표현 할 수 있도록 최대한 유연하게 정의하였다. 에이전트 프로그램과 서버를 이용하여 worm 등의 비정상 트래픽의 시그니처 자동 추출 및 수집을 하였다. 비정상 트래픽의 스캐닝 시도 등을 시그니처 plug-in 으로 정의한 에이전트를 개발하였다. 지금의 연구에서 더 나아가 에이전트 프로그램이 설치된 호스트를 더 늘리고 시그니처의 패턴의 다양화 및 개체 수 확보에 집중 해야 하겠다.

참 고 문 헌


[1] IANA, IANA port number list, http://www.iana.org/assignments/port-numbers/.

[2] A.W. Moore, and K. Papagiannaki, "Toward the Accurate Identification of Network Applications," Passive and Active Measurement Workshop 2005, Boston, MA, USA, Mar. 31-Apr. 1, 2005.

[3] Byung-Chul Park, Young J. Won, Myung-Sup Kim, and James Won-Ki Hong. "Towards Automated Application Signature Generation for Traffic Identification," Proc. of the IEEE/IFIP Network Operations and Management Symposium (NOMS 2008), Salvador, Brazil, Apr. 2008, pp. 160-167.

[4] T.S. Choi, C.H. Kim, S. Yoon, J.S. Park, B.J. Lee, H.H. Kim, H.S. Chung, and T.S. Jeong, "Content-aware Internet Application Traffic Measurement and Analysis," Proc. of the IEEE/IFIP Network Operations and Management Symposium (NOMS 2004), Seoul, Korea, Apr. 23, 2004, Vol. 1, pp. 511-524.

[5] T. Karagiannis, K. papagiannaki, and M. Faloutsos, "BLINC: Multilevel Traffic Classification in the Dark," ACM SIGCOMM 2005," Philadelphia, PA, USA, Aug. 21-26, 2005.

[6] T. Karagiannis, A. Broido, M. Faloutsos, and Kc claffy, "Transport Layer Identification of P2P Traffic," Internet Measurement Conference, Proc. of the 4th ACM SIGCOMM conferences on Internet measurement, Taormina, Sicily, Italy, Oct. 25-27, 2004, pp. 121-134.