한국통신학회논문지 '13-05 Vol.38B No.05
http://dx.doi.org/10.7840/kics.2013.38B.5.368

인터넷 응용 트래픽 분석을 위한 행위기반 시그니쳐 추출 방법

윤 성 호◆, 김 명 섭°

Behavior Based Signature Extraction Method for Internet Application Traffic Identification

Sung-Ho Yoon◆, Myung-Sup Kim°

요 약

최근 급격한 인터넷의 발전으로 효율적인 네트워크관리를 위해 응용 트래픽 분석의 중요성이 강조되고 있다. 본 논문에서는 기존 분석 방법의 한계점을 보완하기 위하여 행위기반 시그니쳐를 이용한 응용 트래픽 분석 방법을 제안한다. 행위기반 시그니쳐는 기존에 제안된 다양한 트래픽 특징을 조합하여 사용할 뿐만 아니라, 복수 개 플로우들의 첫 질의 패킷을 분석 단위로 사용한다. 제안한 행위기반 시그니쳐의 타당성을 검증하기 위해 국내외 응용 5종을 대상으로 정확도를 측정결과, 모든 응용에서 100% Precision을 나타내었다.

Key Words : behavior based signature, signature extraction, signature creation, traffic identification, traffic classification

ABSTRACT

The importance of application traffic identification is emphasized for the efficient network management with recent rapid development of internet. In this paper, we present the application traffic identification method using the behavior based signature to improve the previous limitations. The behavior based signature is made by combining the existing various traffic features, and uses the Inter-Flow unit that is combination of the first request packet of each flow. All signatures have 100% precision when measured the accuracy of 5 applications using at home and abroad to prove the feasibility of the proposed signature.

Ⅰ. 서 론

초고속 인터넷의 보급과 인터넷 기반의 서비스가 다양화됨에 따라 네트워크 관리의 중요성이 강조되고 있다. 네트워크 이용자(end user) 측면에서는 고품질 서비스의 안정적인 제공에 대한 요구가 증대되고, 사업자(ISP: Internet Service Provider, ICP: Internet Contents Provider) 측면에서는 망 관리 비용을 최소화하면서 다양한 고품질의 서비스를 제공하기 위한 요구가 증대되고 있다[1,2]. 하지만 한정적인 네트워크 자원과 급증하는 트래픽은 네트워크의 부담을 가중시킨다. 또한, 특정 네트워크 구간에서의 사용자 급증과 시간대별 병목 현상은 통신 속도를 급격히 저하시킨다. ISP나 네트워크 관리자는 망

※ 이 논문은 정부(교육과학기술부)의 재원으로 2010년도 한국연구재단-차세대정보컴퓨팅기술개발사업(20100020728) 및 2012년도 한국연구재단(2012R1A1A2007483)의 지원을 받아 수행된 연구임.
◆ 주저자 : 고려대학교 컴퓨터정보학과 네트워크 관리 연구실, sungho_yoon@korea.ac.kr, 학생회원
° 교신저자 : 고려대학교 컴퓨터정보학과 네트워크 관리 연구실, tmskim@korea.ac.kr, 종신회원
논문번호 : KICS2013-02-093, 접수일자 : 2013년 2월 14일, 최종논문접수일자 : 2013년 4월 26일

의 안정성과 신뢰성을 확보하기 위해 네트워크 장비의 대역폭을 증가시키고 확장하는 방법을 취하지만, 무작정 네트워크 장비의 확충과 성능 향상을 고집하기에는 비용과 기술적인 측면에서 무리가 있다.

효과적인 네트워크 자원 활용을 위해 응용 레벨 트래픽 분석을 기반으로 사전에 네트워크 소모량이 높은 서비스와 피크 시간대 등을 파악하고 서비스의 사용자 별 이용 패턴을 분석하여 모니터링 해야 한다. 정확한 트래픽 분석을 통한 Traffic Engineering, Network Planning, QoS(Quality of Service) Planning, SLA(Service Level Agreement), Billing 등의 다양한 네트워크 관리 정책의 수립이 요구되는 시점이다.

네트워크 트래픽의 응용을 탐지하는 트래픽 분석은 다양한 네트워크 관리 정책들을 적용하기 위해서 반드시 필요한 선행 기술이다. 트래픽 분석 방법론 또는 시스템의 최종목표는 분석하고자 하는 대상 네트워크의 모든 트래픽을 응용 별로 정확하게 분석하는 것이다.

트래픽 분석을 위해 다양한 트래픽 특징을 이용한 방법론들이 제안되었지만, 실제 네트워크 관리에 활용하기에는 많은 한계점을 가지고 있다. 대표적인 한계점으로는 동적 또는 임의 포트 사용, 시그니쳐 생성 및 관리, 계산 복잡도, 사생활 침해, 실시간 제어 문제 등이 있다.

본 논문에서는 기존 분석 방법의 한계점을 보완하기 위하여 행위기반 시그니쳐를 제안한다. 대부분의 인터넷 응용들은 특정 기능(로그인, 파일 전송, 채팅 등)을 사용할 때, 2개 이상의 플로우를 발생시킨다. 이때 발생되는 플로우에서 추출된 특징들은 다른 응용과 구별되는 패턴을 가지고 있다. 따라서 본 논문에서는 이러한 패턴을 이용하여 행위기반 시그니쳐를 제안한다. 본 시그니쳐는 기존에 제안된 다양한 트래픽 특징을 조합하여 사용할 뿐만 아니라, 복수 플로우의 특정 패킷들을 조합한 플로우 간(inter-flow) 단위를 사용함으로써 특정 응용을 매우 정확하게 분석할 수 있다. 즉, 특정 응용을 사용할 때 발생하는 복수 개 플로우들의 첫 질의 패킷에서 다양한 트래픽 특징(목적지 IP, 목적지 포트 번호, 첫 N 바이트 페이로드 등)을 엔트리로 추출하고 이를 일련의 순서 또는 임의의 순서로 조합하여 행위기반 시그니쳐를 생성함으로 계산 복잡도, 사생활 침해, 실시간 제어 문제를 해결한다.

본 논문의 주요 내용은 다음과 같다. 첫째, 새로운 분석 단위, 플로우 간(inter-flow) 단위를 제안한다. 응용을 사용할 때 발생하는 복수 플로우의 특정 패킷들을 조합하여 사용함으로써 실시간 제어가 가능한 패킷 단위 분석의 장점과 다양한 트래픽 특징을 활용할 수 있는 플로우 단위 분석의 장점을 모두 가질 수 있다. 특히, 여러 패킷들의 특징을 조합하여 사용함으로 추출의 범위가 패킷 또는 플로우 단독으로 사용하는 방법론에 비해 시그니쳐 생성이 용이하며 추출된 시그니쳐는 매우 정확하게 해당 응용 트래픽을 분석한다.

둘째, 행위기반 시그니쳐 모델을 제시하고, 이를 기반으로 첫 질의 패킷 추출 모듈, 후보 시그니쳐 추출 모듈, 시그니쳐 선택 모듈, 총 3단계의 시그니쳐 추출 알고리즘을 제안한다. 특히 후보 시그니쳐 모듈과 시그니쳐 선택 모듈에서는 다양한 임계값들을 사용하여 시그니쳐 추출의 생산성과 정확도를 향상시킨다.

본 논문은 다음과 같은 순서로 기술한다. 2장에서는 기존 트래픽 분류 방법론들에 대해 살펴보고, 3장에서는 행위기반 시그니쳐를 정의한다. 4장에서 시그니쳐 추출 알고리즘을 제시하고 5장에서는 제안한 행위기반 시그니쳐의 타당성을 증명하기 위한 실험 결과를 기술한다. 마지막으로 6장에서는 결론과 향후 연구를 언급한다.

Ⅱ. 관련 연구

트래픽 분석 방법론은 그 중요성이 증가함에 따라 지속적으로 연구가 진행되고 있다. 트래픽 분석 방법들은 트래픽 분석 시 사용하는 트래픽 특징을 기준으로 포트기반 분석[3,4], 페이로드기반 분석[5,6], 통계정보기반 분석[7,8], 상관관계기반 분석[9] 등으로 구분된다. 또한, 분석 단위 기준으로 패킷기반 분석, 플로우기반 분석으로 구분될 수 있다.

포트기반 분석은 Internet Assigned Number Authority (IANA)[3]에서 지정한 포트 정보를 이용한다. 포트 번호와 대응하는 서비스(HTTP(80), telnet(23), e-mail(25,110), FTP(20,21))를 기준으로 분석하기 때문에 적은 메모리 사용으로 매우 빠르게 분석할 수 있는 장점을 가진다. 하지만, 최근 사용되는 응용들은 방화벽 및 IPS 장비를 통과하기 위해 포트 번호를 임의로 설정하여 트래픽을 발생시키므로 더 이상 포트 번호가 특정 서비스, 프로토콜을 의미하지 않는다. 또한, Torrent, Skype와 같은 응용에서는 포트 번호를 사용자가 설정하거나 매 실행 시 임의의 포트 번호를 사용하기도 한다.

이러한 문제를 해결하기 위해 패킷의 페이로드 내에서 응용마다 가지는 특정 스트링(시그니쳐)의 포함 유무를 통해 트래픽을 분석하는 페이로드기반 분석 방법이 제안되었다. 트래픽의 내용을 확인하기 때문에 분석 성능(분석률, 정확도)이 매우 높지만, 시그니쳐 생성 및 관리, 암호화 트래픽, 높은 계산 복잡도, 패킷 단편화, 사생활 침해 등과 같은 많은 한계점을 가지고 있다[10].

트래픽 암호화 및 사생활 침해 문제를 해결하기 위해 트래픽 내용을 보지 않고 패킷 및 윈도우 크기, 패킷 간 시간 간격 등과 같은 통계적 특징만을 이용한 통계 기반 분석 방법이 제안되었다. 이 방법론은 패킷의 헤더 정보를 통해 통계 정보를 생성하므로 기존 트래픽 분류 방법론들의 한계점들을 보완할 수 있다. 하지만, 같은 엔진 기반의 응용이거나 같은 응용 레벨 프로토콜을 사용하는 경우 동일한 통계적 특징을 가지기 때문에 상세한 응용 별 분석이 어려운 한계점을 가진다.

최근에는 전통적인 트래픽 분석 방법의 한계점을 보완하기 위해 패킷 단위의 트래픽을 플로우 단위로 변경하고 이들의 상관관계를 분석하는 방법이 제안되었다[7-9]. 플로우는 5-tuple(SrcIP, SrcPort, DstIP, DstPort, Transport Layer Protocol)이 동일한 패킷의 집합을 의미한다. 플로우의 크기, 기간 등과 같은 통계 정보와 플로우들 간의 연결 형태를 이용하여 트래픽을 분석한다. 패킷기반 분석 방법 보다 다양한 특징을 사용할 수 있기 때문에 다양한 분석이 가능하지만, 플로우 생성이 완료 될 때까지 분석하지 못하며, 플로우의 통계 정보를 계산하는 오버헤드가 발생한다. 또한, 유사한 통계 정보를 가지는 응용 간 구별이 어려운 문제점을 가지고 있다.

분석 단위 기준으로 트래픽 분석 방법론을 구분해 보면 패킷 단위와 플로우 단위로 구분 할 수 있다. 패킷 단위의 트래픽 분석은 해당 패킷 내에서 특정 응용을 구분할 수 있는 트래픽 정보(헤더 정보나 비트 스트링)를 시그니쳐로 추출하여 트래픽을 분석한다. 실시간으로 발생하는 패킷을 기반으로 분석하기 때문에 실시간 분석이 가능하다는 장점을 가진다. 하지만, 추출 대상 범위가 단일 패킷 내로 제한적이기 때문에 특정 응용을 대표하는 시그니쳐를 추출하는 것은 매우 어렵다. 또한, 트래픽 분석 시 모든 패킷의 헤더 정보와 페이로드 정보를 분석해야 하는 오버헤드가 발생한다.

플로우 단위 트래픽 분석은 플로우를 구성하는 복수 패킷들의 통계 정보와 연결 형태, 발생 시점 등을 이용하여 트래픽을 분석하기 때문에 패킷 단위보다 다양한 트래픽 특징을 사용할 수 있다. 따라서 시그니쳐 생성과정이 비교적 용이하지만, 비슷한 통계적 특징을 가지는(동일 엔진 및 프로토콜 사용) 응용의 트래픽이 많아짐에 따라 응용 별 트래픽 분석 정확도가 매우 낮고, 플로우가 완성되는 시점에 분석이 이루어짐으로 실시간 제어에 어려움이 있다.

최근까지 제안된 분석 방법들은 기존의 한계점을 극복하기 위해 점점 발전하고 있지만 여전히 많은 한계점을 가진다. 이를 해결하기 위해 본 논문에서는 다양한 트래픽 특징(목적지 IP, 목적지 포트 번호, 첫 N 바이트 페이로드 등) 조합과 새로운 분석 단위(복수 플로우의 첫 질의 패킷들)를 사용한다. 즉, 페이로드의 첫 N 바이트만을 사용하여 계산 복잡도 문제와 사생활 침해 문제를 해결하고, 플로우의 첫 패킷만을 사용하여 실시간 제어가 가능하도록 한다. 또한, 여러 패킷들의 특징을 조합함으로써 특정 응용을 대표하는 시그니쳐 생성을 용이하게 한다.

Ⅲ. 행위기반 시그니쳐

본 장에서는 행위기반 시그니쳐를 정의하기 위해 시그니쳐의 속성으로 사용하는 트래픽의 특징과 트래픽 분석 단위를 설명한다. 또한 행위기반 시그니쳐 모델을 제시하고 국내에서 많이 사용되는 응용 (Nateon)을 선정하여 시그니쳐 실제 예시를 보인다.

행위 기반 시그니쳐에서 사용하는 트래픽 특징은 총 4가지 이다. 목적지 IP, 목적지 포트 번호, 전송 계층 프로토콜, 첫 N 바이트 페이로드이다. 트래픽의 헤더 정보(IP, 포트, 프로토콜)는 해당 응용이 서버-클라이언트 연결을 사용하거나 고정 포트를 사용하는 경우 큰 의미를 가진다. 페이로드 정보는 응용을 식별하는 중요한 키를 가지고 있지만 최근 사생활 침해 문제와 계산 복잡도 문제로 인해 사용을 꺼리고 있다. 이를 해결하기 위해 행위기반 시그니쳐는 첫 N 바이트만을 사용한다. HTTP트래픽인 경우 Method 키워드만 추출되는 것을 방지하기 위하여 N의 값을 10이상으로 설정하고 Non-HTTP 트래픽은 2이상으로 설정한다. 전체가 아닌 일부 페이로드만을 사용함으로써 사생활 문제를 해결할 뿐만 아니라 고정된 위치(offset, length)의 페이로드를 사용하기 때문에 계산 복잡도 문제도 해결할 수 있다.

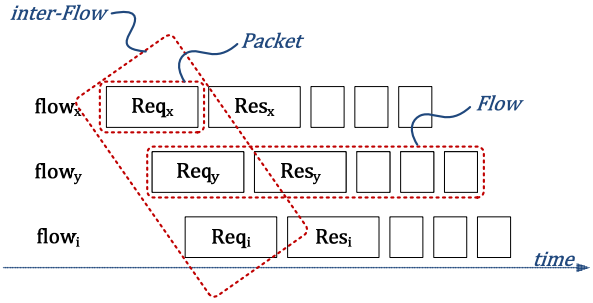

그림 1. 트래픽 분석 단위
Fig. 1. Units of Traffic Identification

그림 1은 트래픽 분석 시 대상이 되는 트래픽의 다양한 단위를 보여준다. 본 논문에서는 패킷 단위, 플로우 단위 트래픽 분석의 문제점을 보완하고 각 단위의 장점을 활용하기 위해 플로우 간(inter-flow) 단위를 사용한다. 복수개의 플로우를 대상으로 시그니처를 생성하기 때문에 시그니처 생성 범위가 넓고, 특정 위치(플로우의 첫 패킷)를 검사하기 때문에 실시간 제어가 가능하다. 즉, 단일 패킷, 단일 플로우를 대상으로 시그니처를 적용하는 것이 아닌 여러 플로우를 대상으로 시그니처를 적용한다. 특히 플로우의 첫 번째 패킷, 질의 패킷들을 대상으로 시그니처를 적용함으로써 단독으로 사용할 수 없는, 단순한 트래픽 특징을 조합하여 시그니처로 사용할 수 있을 뿐만 아니라 단일 패킷, 단일 플로우에 적용하는 방법보다 정확하게 트래픽을 분석할 수 있다.

행위기반 시그니처는 엔트리(Entry)의 조합으로 구성되며 각각의 엔트리는 트래픽의 특징을 가진다. 수식 1, 2는 각각 행위기반 시그니처와 행위기반 시그니처를 구성하는 엔트리를 나타낸다.

$$BS = \left\{ \begin{array}{l} A, T, I, E_1, E_2, \ldots E_n \mid \\ n \geq 2, \\ Src(E_1) = Src(E_2) = \ldots = Src(E_n) \end{array} \right\} \quad (1)$$

$$E = \{X \mid X \subset \{ip, port, prot, payload\}, X \neq \varnothing\} \quad (2)$$

행위기반 시그니처(BS)는 응용이름(A), 타입(T), 인터벌(I), 2개 이상의 엔트리(E)로 구성되며, 엔트리는 목적지 IP(ip), 목적지 포트 번호(port), 전송계층 프로토콜(prot), 그리고 첫 N 바이트 페이로드(payload)로 구성되는 집합의 멱집합(power set)으로 구성되며 공집합은 제외된다. 즉, 응용의 특성상 특정 속성이 의미가 없는 경우, 의미 있는 속성만 선택하여 사용한다. 예를 들어 특정 응용이 P2P 연결 형태와 임의 포트 번호를 사용하는 경우 목적지 IP와 목적지 포트 번호는 의미가 없기 때문에 엔트리의 원소에서 제외한다. 행위 시그니처는 특정 호스트를 기준으로 추출, 적용되기 때문에 모든 엔트리의 출발지 IP는 동일하여야 한다. 표1은 행위기반 시그니처의 각 속성에 대한 설명을 나타낸다.

표 1. 행위기반 시그니처 속성 및 설명
Table 1. Attribute and Explanation of Behavior Signature

Attribute		Explanation
A		Application Nate
T		Applying Type Seq(Sequence), Set(Set)
I		Interval Applying All Entries (ms)
E	ip	Destination IP Address (CIDR)
	port	Destination Port Number
	prot	L4 Protocol (TCP, UDP)
	payload	First N Bytes Payload (HTTP: more than 10bytes Non-HTTP: more than 2bytes)
Src(E_x)		Source IP Address of Entry x

응용 이름(A)은 해당 시그니처로 분석된 트래픽에 분석 결과를 명명하기 위해 기술된다. 타입(T)은 Seq(Sequence)와 Set타입이 있다. Seq는 엔트리들의 순서와 복수 플로우에서 추출한 엔트리가 정확하게 일치되는 것을 의미하고 Set은 순서에 상관없이 일정 인터벌 이내에 모든 엔트리가 일치되는 것을 의미한다. 인터벌(I)은 첫 엔트리와 마지막 엔트리가 매칭되는 일정한 시간 간격(ms)을 의미한다. 즉, 트래픽 발생 시간을 기준으로 해당 패턴이 적용되는 기간을 의미한다.

엔트리(E)은 목적지 IP(ip), 목적지 포트 번호(port), 전송 계층 프로토콜(prot), 첫 N 바이트 페이로드(payload)로 구성된다. 목적지 IP와 포트 번호는 는 해당 엔트리가 전송되는 목적지 IP 주소와 포트 번호를 의미하며, IP의 경우 CIDR 표기법을 이용하여 표기한다. 전송 계층 프로토콜은 해당 엔트리가 전송될 때 사용되는 전송 계층 프로토콜(TCP, UDP)를 의미한다. 페이로드 전체를 엔트리 구성 요소로 사용하지 않고, 페이로드의 최소 첫 N 바이트만 사용한다. HTTP를 사용하는 트래픽의 경우 트래픽의 첫 부분에 위치하는 Method(GET, POST, PUT 등)를 구별하기 위해

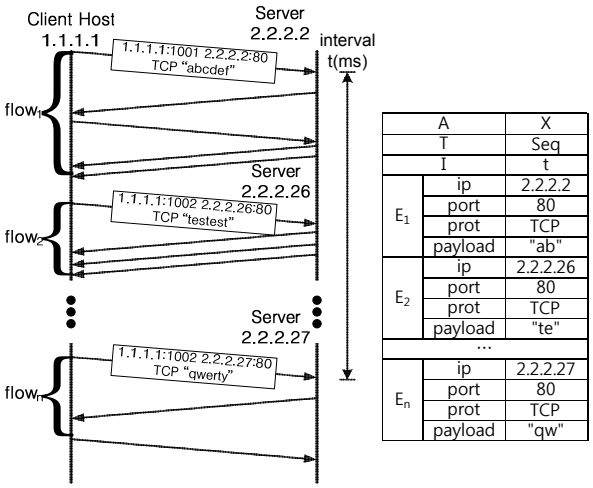

그림 2. 시그니처 예시
Fig. 2. Example of Signature

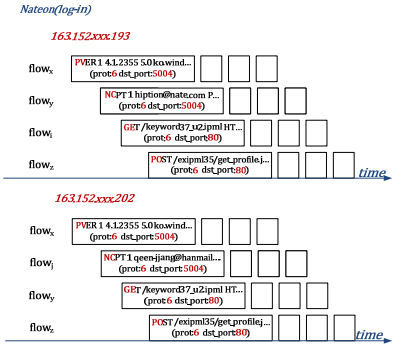

그림 3. Nateon 응용 로그인 시 발생 트래픽
Fig. 3. Nateon Login Traffic

페이로드의 첫 10 바이트 이상을 사용하고 Non-HTTP인 경우, 페이로드의 첫 2바이트 이상을 사용한다.

그림 2는 앞서 제안한 시그니처의 예를 나타낸다. 특정 응용(X)를 사용할 때, 인터벌 t(ms)이내에 발생하는 플로우가 n개인 경우, 해당 시그니처는 n 개의 엔트리를 가진다. 각각의 엔트리들은 각 플로우의 첫 질의 패킷의 트래픽 특징들을 가지고 있다.

앞서 정의한 행위기반 시그니처의 적용 가능성 여부를 파악하기 위해 국내에서 많이 사용하는 응용 2종(Nateon, Skype)을 선정하여 로그인시 발생하는 트래픽을 관찰하였다.

그림 3은 Nateon 응용이 로그인할 때 발생하는 플로우들의 질의 패킷을 보인다. 서로 다른 두 개의 호스트에 발생하는 트래픽 이지만, 동일한 특징(목적지 포트 번호, 전송계층 프로토콜, 페이로드의 첫 2 바이트)을 보여준다. Nateon의 경우 로그인을 실행할 때

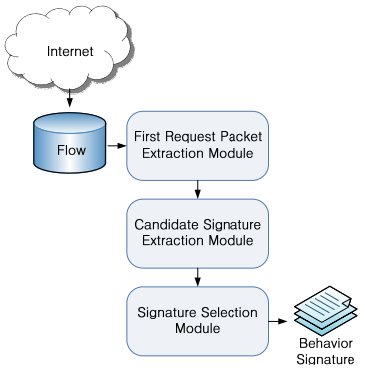

그림 4. 행위기반 시그니처 추출 알고리즘
Fig. 4. Extraction Algorithm of Behavior Signature

고정된 순서의 질의 패킷이 발생되는 것을 확인하였다.

Ⅳ. 추출 알고리즘

본 장에서는 행위기반 시그니처 추출 알고리즘을 첫 질의 패킷 추출 모듈, 후보 시그니처 추출 모듈, 그리고 시그니처 선택 모듈로 구분하여 각각의 모듈에 대한 알고리즘을 기술한다. 그림 4는 각 세부 모듈과 입출력 데이터를 보여준다.

최초, 입력 받은 트래픽에서 첫 질의 패킷에서 엔트리를 추출하여 리스트 형태로 구성하고, 해당 리스트에서 모든 엔트리 조합을 후보 시그니처로 추출한다. 추출된 후보 시그니처 중에서 2대 이상의 호스트에서 공통으로 발생된 후보 시그니처를 행위기반 시그니처를 추출한다.

4.1. 첫 질의 패킷 추출 모듈

본 절에서는 첫 질의 패킷 추출 모듈에 대해 기술한다. 본 모듈은 플로우 단위로 구분된 패킷들을 입력받아 각 플로우의 첫 질의 패킷에서 행위 시그니처 모델에서 정의한 엔트리를 추출하여 리스트로 구성한다. 그림 5는 첫 질의 패킷 추출 모듈의 입출력 데이터를 나타낸다.

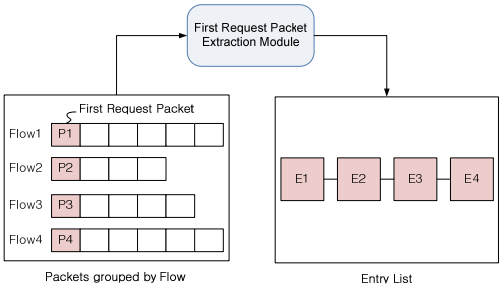

그림 5. 첫 질의 패킷 추출 모듈 입출력 데이터
Fig. 5. Input-Output Data of First Request Packet Extraction Module

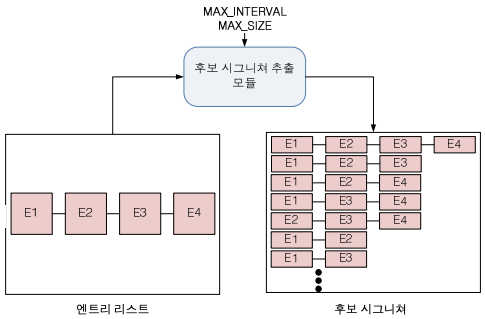

그림 6. 후보 시그니처 추출 모듈 입출력 데이터
Fig. 6. Input-Output Data of Candidate Signature Extraction Module

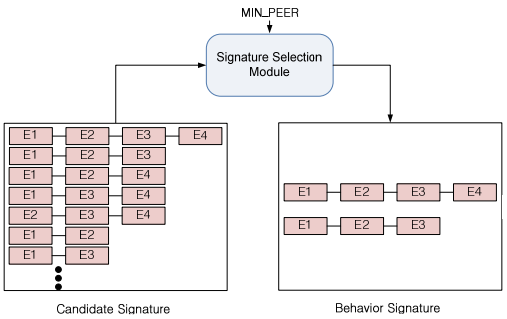

그림 7. 시그니처 선택 모듈 입출력 데이터
Fig. 7. Input-Output Data of Signature Selection Module

입력 받은 모든 플로우에서 첫 질의 패킷을 통해 엔트리 리스트를 구성하고 시간의 순서로 정렬한다. 정렬된 리스트는 본 모듈의 출력으로 다음 모듈에 전달된다.

4.2. 후보 시그니처 추출 모듈

본 절에서는 후보 시그니처 추출 모듈에 대해 기술한다. 본 모듈은 앞서 첫 질의 패킷 추출 모듈의 출력인 엔트리 리스트를 입력 받아 추출 가능한 모든 후보 패턴을 생성한다. 그림 6은 후보 패턴 추출 모듈의 입출력 데이터를 보여준다.

모든 가능 후보 시그니처를 추출하는 것은 매우 높은 계산 복잡도를 가지기 때문에 최대 인터벌(MAX_INTERVAL)과 최대 엔트리 개수(MAX_SIZE)를 임계값으로 설정하여 해당 인터벌 이내에 최대 엔트리 개수 이내로 후보 시그니처를 추출한다. 즉, 입력 받은 엔트리 리스트의 첫 번째 엔트리를 시작으로 최대 인터벌(MAX_INTERVAL) 크기만큼 구간을 설정하고 해당 구간의 엔트리들을 대상으로 최대 엔트리 개수(MAX_SIZE)이내의 추출 가능한 모든 후보 시그니처를 추출한다. 후보 시그니처는 본 모듈의 출력으로 다음 모듈에 전달된다.

4.3. 시그니처 선택 모듈

본 절에서는 시그니처 선택 모듈에 대해 기술한다. 앞서 추출된 후보 패턴 중에서 최소 호스트 개수(MIN_PEER)을 초과한 패턴들에 한해 시그니처로 선택한다. 그림 7은 시그니처 선택 모듈의 입출력 데이터를 보여준다.

행위 기반 시그니처는 특정 호스트에 종속되지 않고 모든 호스트에서 특정 응용을 사용할 때 공통으로 발생하는 패턴을 의미한다. 본 모듈에서 사용하는 임계값(MIN_PEER)은 2이상의 값으로 설정한다.

즉, 2대 이상의 호스트에서 특정 응용을 사용할 때 공통으로 발생한 후보 시그니처를 선택한다. 임계값이 증가하면, 시그니처의 정확도는 증가하지만, 시그니처 개수가 감소하여 분석률이 감소하므로 정답지 생성 시 사용된 호스트의 개수를 감안하여 적절한 임계값을 설정한다. 또한, 적은 개수의 엔트리 개수를 가지는 시그니처가 엔트리 개수가 많은 시그니처에 포함될 가능성이 있기 때문에 추출된 시그니처의 포함관계를 확인하여 포함관계가 있는 경우에는 시그니처 선택에서 제외한다.

V. 실 험

본 장에서는 행위기반 시그니처의 타당성을 증명하기 위해 국내외 응용 5종을 선정하여 시그니처를 추출하고 평가한 결과를 기술한다.

5.1. 실험 환경

국내외에서 많은 사람들이 사용하는 응용 5종(Nateon: 메신저, DropBox: 웹저장소, UTorrent: P2P 파일 전송, Skype: 메신저, Teamviewer: 원격제어)을 선정하였다. 정확한 성능 평가를 위해 4대의 서로 다른 호스트에서 다른 날짜에 2회에 걸쳐 응용 트래픽을 수집하였다. 특정 응용의 트래픽을 정확하게 수집하기 위해 수집 대상 호스트에 소켓 정보를 수집하는 에이전트를 설치하여 트래픽을 수집하였다[11]. 본 실험에서 사용한 임계값은 MAX_INTERVAL 5000ms, MAX_SIZE 10, MIN_PEER는 4로 설정하였다. 표 2는 실험에 사용한 응용 트래픽의 정보를 나타낸다.

5.2. 시그니처 추출 및 정확도 측정

본 논문에서 제안하는 알고리즘을 통해 추출된 시그니처는 표 3과 같다.

표 2. 응용 트래픽 정보
Table 2. Traffic Trace Information

Application	Trace Name	Start	Duration (min)	Host	Flow	Packet	Byte(K)	Function
Nateon	nateon_#01(NA1)	2012:11:16 02:46	17	2	589	224,334	204,069	login chatting file download file upload
	nateon_#02(NA2)	2012:11:22 14:10	6	2	152	58,389	50,040	
DropBox	dropbox_#01(DB1)	2012:11:08 03:53	19	2	231	42,241	34,362	login file download file upload
	dropbox_#02(DB2)	2012:11:22 14:56	2	2	16	1,497	1,346	
UTorrent	utorrent_#01(UT1)	2012:11:08 01:35	23	2	16,573	3,879,615	3,913,652	file download file upload
	utorrent_#02(UT2)	2012:11:22 14:21	3	2	1,546	259,074	268,789	
Skype	skype_#01(SK1)	2012:11:13 01:45	49	2	1,862	158,859	98,736	login chatting file download file upload
	skype_#02(SK2)	2012:11:22 14:26	4	2	350	8,972	4,606	
Teamviewer	teamviewer_#01(TV1)	2012:11:15 11:25	24	2	339	238,562	201,094	login remote access file download file upload
	teamviewer_#02(TV2)	2012:11:22 14:41	4	2	47	19,361	14,750	

Nateon의 경우, 총 48개의 시그니처가 추출되었다. 로그인 시 다른 응용에 비해 다양한 서버(인증서버, 업데이트 서버, pop-up 서버, 메인 서버 등)와 통신하는 구조로 인해 실험에서 설정한 최대 엔트리 개수(10)보다 많은 플로우를 패턴으로 가지는 경우에서 해당 패턴의 모든 부분집합(subset)이 시그니처로 추출되었다. 최대 엔트리 개수 임계값은 시그니처 생산성(추출시간) 측면과 관련 있기 때문에 최적의 임계값을 설정하여 시그니처 개수를 줄이는 연구가 필요하다. 표 3에 기술된 Nateon시그니처 예시는 총 10개의 엔트리들로 구성되어있다. Nateon 응용은 서버-클라이언트 형태로 동작하고 고정 포트 번호를 사용하기 때문에 모든 속성을 사용하였다. 제시한 시그니처 예시는 특정 인터벌(4324ms)이내에 10개의 엔트리를 각각 첫 질의 패킷으로 가지는 플로우들이 순차적(Seq)으로 발생하는 경우 해당 모든 플로우를 Nateon으로 분석한다.

UTorrent의 경우, 총 7개의 시그니처가 추출되었으며, P2P 형태와 임의 포트 번호를 사용하기 때문에 목적지 IP와 목적지 포트 번호를 "any"로 표기했다. 제시한 시그니처 예시는 특정 인터벌(5000ms)이내에 2개의 엔트리가 각각 첫 질의 패킷으로 가지는 플로우에 임의(Set)의 순서로 발생하는 경우 모든 해당 모든 플로우를 UTorrent로 분석한다.

시그니처 추출 대상 트래픽이 특정 날짜 및 특정 버전을 사용할 경우, 날짜 정보와 버전 정보가 엔트이의 Payload 속성값에 포함되는 경우가 있다. Nateon 시그니처의 첫 엔트리 "1.4.1.2485", Skype 시그니처의 첫 엔트리 "5.10"은 클라이언트 프로그램의 버전을 나타낸다. 날짜 정보와 버전 정보가 시그니처에 포함되면 해당 날짜가 아니거나 버전이 아닌 경우 분석되지 않는다. 따라서 시그니처에 포함된 변동 가능한 payload 정보는 관리자의 의해 삭제되어야 하며, 해당 정보를 자동으로 인지하고 삭제하는 모듈은 좀 더 자세한 연구가 필요하다.

해당 시그니처의 정확도를 측정하기 위해 5종 트래픽을 혼합하여 검증 트래픽을 구성하고 개별 응용 별로 정확도(Precision, Recall)를 측정하였다. 정확도를 측정하는 수식은 다음과 같다.

$$\mathrm{Preciaion} = \frac{TP}{TP+FP} \tag{3}$$

$$Recall = \frac{TP}{TP+FN} \tag{4}$$

TP(True Positive)는 특정 응용 X의 시그니처가 해당 응용 X를 정확하게 분석된 양을 의미한다. FP(False Positive)는 X의 시그니처가X가 아닌 응용을 X라 분석한 양을 의미하고, FN(False Negative)는 X의 시그니처가 X를 X가 아니라고 분석한 양을 의미한다. 즉, Precision은 해당 응용으로 분석된 트래픽 중에 정확하게 분석된 비율을 의미하고, Recall은 해당 응용의 전체 트래픽 중에 정확하게 분석된 비율을 의미한다. 표4는 응용 5종의 행위기반 시그니처 정확도를 보여준다.

추출된 모든 시그니처는 정확하게 해당 응용을 분석하였다. 즉, 모든 응용 별 Precision의 값은 1.00이었다. Recall의 경우, 플로우 단위 평균 0.18, 바이트

표 3. 추출 시그니처 개수 및 예시
Table 3. Number and Example of Extracted Signature

Application	Num. of Signature	Example
Nateon	48	{Nateon, Seq, 4324, (203.xxx.xxx.91/32, 5004, 6, "PVER 1 4.1.2485 5.0"), (120.xxx.xxx.0/24, 5004, 6, "NCPT 1"), (117.xxx.xxx.17/32, 80, 6, "GET /keyword37_u2.op"), (203.xxx.xxx.117/32, 80, 6, "POST /client/club/Ge"), (211.xxx.xxx.0/24, 80, 6, "GET /upload/notice/"), (211.xxx.xxx.0/24, 80, 6, "GET /upload/"), (211.xxx.xxx.0/24, 80, 6, "GET /upload/"), (211.xxx.xxx.0/24, 80, 6, "GET /upload/"), (117.xxx.xxx.12/32", 80, 6, "GET /nateon/tickerH"), (120.xxx.xxx.20/32, 80, 6, "POST /client/CountMe")}
DropBox	1	{DropBox, Seq, 3258, (any, 443, 6, "0x16 0x03 0x01 0x00 0x5B 0x01 0x00 0x00 0x57 0x03 0x01 0x50"), (any, 80, 6, "GET /subcribe?host_")}
UTorrent	7	{UTorrent, Set, 5000, (any, any, 17, "d1:ad2:id20:"), (any, any, 17, "A."), (any, any, 17, "d1:ad2:id20:")}
Skype	3	{Skype, Seq, 5000, (any, any, 6, "GET /ui/0/5.10."), (any, any, 6, "0x16 0x03 0x01 0x00")}
Teamviewer	1	{Teamviewer, Seq, 4991, (any, 5938, 6, ".$"), (any, 5938, 17, "0x00 0x00 0x00 0x00 0x00 0x00 0x00 0x00")}

표 4. 행위기반 시그니처의 정확도 측정 결과
Table 4. Accuracy of Behavior Signature

Application	Unit	Precision	Recall
Nateon	flow	1.00 (447/447)	0.60 (447/741)
	byte(K)	1.00 (5,064/5,064)	0.02 (5,064/254,110)
DropBox	flow	1.00 (193/193)	0.78 (193/247)
	byte(K)	1.00 (5,303/5,303)	0.15 (5,303/35,708)
UTorrent	flow	1.00 (2,999/2,999)	0.17 (2,999/18,106)
	byte(K)	1.00 (2,741,745/2,741,745)	0.66 (2,741,745/4,182,441)
Skype	flow	1.00 (127/127)	0.06 (127/2,088)
	byte(K)	1.00 (1,589/1,589)	0.02 (1,589/103,342)
Teamviewer	flow	1.00 (239/239)	0.63 (239/385)
	byte(K)	1.00 (8,237/8,237)	0.04 (8,237/215,845)
Total	**flow**	**1.00 (4,005/4,005)**	**0.18 (4,005/21,487)**
	byte(K)	**1.00 (2,761,938/2,761,938)**	**0.57 (2761938/4,791,446)**

단위 평균 0.57로 응용과 측정 단위에 따라 큰 차이를 보였다. 이는 분석된 트래픽의 통계적 특성 (Heavy 또는 Light 플로우)이 응용 마다 상이하기 때문이다. 여러 호스트에서 공통으로 발생하는 패턴을 시그니처로 사용하기 때문에 낮은 Recall 값을 가지지만 시그니처의 정확도는 매우 높았다. 따라서, 응용 트래픽 분석(monitoring)측면 보다는 응용 트래픽 탐지 및 제어(detection and control)측면에서 활용이 가능하다. 행위 시그니처가 분석한 트래픽의 통계적 특성은 향후 추가적인 연구가 필요하다.

Ⅵ. 결론 및 향후 연구

응용 트래픽 분석은 다양한 네트워크 관리 정책을 수행하기 위해 반드시 선행되어야 하는 작업이다. 하지만, 네트워크에서 발생하는 트래픽이 복잡 다양해지고 있고, 그에 따라 전통적인 트래픽 분석 방법으로는 모든 트래픽을 분석하기 어려워졌다.

본 논문에서는 복수 플로우의 첫 질의 패킷에서 트래픽 특징을 추출하여 행위기반 시그니처를 추출하는 방법을 제시하였다. 이는 기존 패킷 단위 및 플로우 단위 트래픽 분석의 한계점을 보완한다. 제안한 행위기반 시그니처의 타당성을 증명하기 위해 국내외 응용 5종을 선정하여 시그니처를 추출하고 정확도를 추출하였다. 비록 Recall 측면에서는 낮은 값을 보이는 응용도 존재하였지만, 모든 응용에서 100% Precision을 보였다. 이는 분석된 트래픽은 해당 응용으로 정확하게 분석되었다는 의미이다.

향후 연구로써는 추출된 시그니처가 응용의 어떤 기능을 탐지하는지 확인하는 "Function Naming" 모듈과, 추출된 시그니처에 날짜 정보와 버전 정보를 삭제하는 "Signature Arrangement" 모듈을 추가하고자 한다. 또한, 기존 페이로드, 통계 기반 시그니처와 비교를 통해 행위 시그니처의 타당성을 증명하고 암호화 트래픽의 분석 가능성에 관한 연구를 진행할 계획이다.

References

[1] S.-H. Yoon and M.-S. Kim, "A study of performance improvement of internet application traffic identification using flow correlation," *J. KICS*, vol. 36, no. 6, pp. 600-607, May 2011.

[2] S. Sen and J. Wang, "Analyzing peer-to-peer traffic across large networks," in *Proc. Internet Measurement Conf. (IMC)*, pp. 137-150, Marseille, France, Nov. 2002.

[3] IANA, *IANA port number list*, Retrieved 5, 24, 2013, from http://www.iana.org/assignments/service-names-

port-numbers/service-names-port-numbers.xml.

[4] J. Zhang and A. Moore, "Traffic trace artifacts due to monitoring via port mirroring," in *Proc. End-to-End Monitoring Techniques and Services (E2EMON)*, pp. 1-8, Munich, Germany, May 2007.

[5] F. Risso, M. Baldi, O. Morandi, A. Baldini, and P. Monclus, "Lightweight, payload-based traffic classification: an experimental evaluation," in *Proc. IEEE Int. Conf. Commun (ICC) '08*, pp. 5869-5875, Beijing, China, May 2008.

[6] J.-S. Park, S.-H. Yoon, and M.-S. Kim, "Software architecture for a lightweight payload signature-based traffic classification system," in *Proc. 3rd Int. Conf. Traffic Monitoring and Analysis (TMA) '11,* pp. 136-149, Vienna, Austria, Apr. 2011.

[7] K. Xu, Z.-L. Zhang, and S. Bhattacharya, "Profiling internet backbone traffic: behavior models and applications," in *Proc. ACM SIGCOMM 2005*, pp. 169-180, Philadelphia, U.S.A., Aug. 2005.

[8] A. W. Moore and D. Zuev, "Internet traffic classification using bayesian analysis techniques," in *Proc. ACM SIGMETRICS,* pp. 50-60, Banff, Canada, June 2005.

[9] T. Karagiannis, K. Papagiannaki, and M. Faloutsos, "BLINC: multilevel traffic classification in the dark," in *Proc. ACM SIGCOMM 2005*, pp. 229-240, Philadelphia, U.S.A., Aug. 2005.

[10] A. Callado, C. Kamienski, G. Szabo, B. Gero, J. Kelner, S. Fernandes, and D. Sadok, "A survey on internet traffic identification," *IEEE Commun. Surveys Tutorials*, vol. 11, no. 3, pp. 37-52, July 2009.

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

윤 성 호 (Sung-Ho Yoon)

2009년 고려대학교 컴퓨터 정보학과 졸업

2011년 고려대학교 컴퓨터 정보학과 석사

2011년~현재 고려대학교 컴퓨터 정보학과 박사과정

<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

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