인터넷 트래픽 분석을 위한 행위기반 시그니쳐 생성 방법론 개발과 적용에 관한 연구

윤성호 O, 김명섭

고려대학교 컴퓨터정보학과

{sungho_yoon, tmskim}@korea.ac.kr

A Study of the Development and Application of Behavior-based Signature Creation Method for Internet Traffic Identification

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 request 패킷을 분석 단위로 사용한다. 제안한 행위기반 시그니쳐의 타당성을 검증하기 위해 국내외 응용 5 종을 대상으로 정확도를 측정하고 페이로드기반 시그니쳐와 비교한 결과를 제시한다.

1. 서론

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.

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, 2 개 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.

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

2. 관련연구

트래픽 분석 방법론은 대용량 트래픽의 발생과 다양한 인터넷 응용이 개발됨에 따라 지속적으로 연구가 진행되고 있다. 단순히 포트 번호를 이용한 분석 방법부터 페이로드, 통계정보, 상관관계에 이르기 까지 다양한 트래픽 특징을 이용한 분석 방법

* 이 논문은 정부(교육과학기술부)의 재원으로 2010 년도 한국연구재단-차세대정보컴퓨팅기술개발사업(20100020728) 및 2012 년도 한국연구재단(2012R1A1A2007483)의 지원을 받아 수행된 연구임.

이 제안되었다. 분석 방법론의 발전과 함께 인터넷 응용 개발자들은 자신이 개발한 응용이 인터넷 상에서 원활히 사용될 수 있도록 트래픽의 발생 형태를 점점 복잡 다양하게 하고 있다. 따라서 오늘날에는 단순히 트래픽의 특정 특징만을 사용하는 방법보다는 응용의 고유한 트래픽 발생 행위를 분석하는 행위기반 트래픽 분석 방법론들이 제안되고 있다. 본 장에서는 기존에 제안된 대표적인 행위기반 분석 방법과 한계점에 대해 설명한다.

T. Karagiannis[3]는 페이로드와 포트정보를 사용하지 않고 인터넷 트래픽을 분석하기 위해 호스트에서 발생하는 트래픽의 행위를 3 가지 레벨(social, functional, application)로 구분하였다. 이 방법은 매우 간단하고 사용하기 용이하여 다양한 네트워크에 적용이 가능하지만, 특정 호스트가 단일 응용만 사용한다는 가정과 세밀한 응용 별 분석이 불가능하다는 한계점을 가진다. L. Bernaille[4]는 응용 별 실시간 분석을 위해 초기 K 개 패킷 크기(K-data-packet-size)를 사용하여 트래픽을 분석한다. 페이로드 정보를 사용하지 않고 단순히 해당 패킷의 크기와 방향을 이용하기 때문에 사생활 침해 문제를 해결할 수 있지만, 동일 프로토콜을 사용한 응용의 경우 구별이 힘들다는 한계점을 가지고 있다. 또한, 패킷 크기라는 단일 특징을 사용하기 때문에 여러 응용에서 중복되지 않는 고유한 패턴을 찾기가 쉽지 않다.

본 논문에서는 기존의 행위기반 분석 방법론의 한계점을 극복하기 위해 특정 행위 시 발생하는 플로우들의 첫 request 패킷에서 특징을 시그니쳐로 추출하여 트래픽 분석에 사용한다. 복수개의 플로우에서 특징을 찾기 때문에 응용 별로 중복되지 않는 고유한 시그니쳐 생성이 용이하다. 또한, 고정된 위치의 페이로드 일부만을 사용함으로써, 계산 복잡도 및 사생활 침해 문제를 해결한다.

3. 행위기반 시그니쳐

본 장에서는 행위기반 시그니쳐를 정의하기 위해 시그니쳐의 속성으로 사용하는 트래픽의 특징들과 트래픽 분석 단위를 설명한다. 행위 기반 시그니쳐에서 사용하는 트래픽 특징은 총 4 가지 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. 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 은 트래픽 분석 시 대상이 되는 트래픽의 다양한 단위를 보여준다. 본 논문에서는 패킷 단위, 플로우 단위 트래픽 분석의 한계점을 보완하고 각 단위의 장점을 활용하기 위해 플로우 간(inter-flow) 단위를 사용한다. 복수개의 플로우를 대상으로 시그니쳐를 생성하기 때문에 시그니쳐 생성 범위가 넓고 특정 위치(플로우의 첫 패킷)를 검사하기 때문에 실시간 제어가 가능하다. 즉, 단일 패킷, 단일 플로우를 대상으로 시그니쳐를 적용하는 것이 아닌 여러 플로우를 대상으로 시그니쳐를 적용한다. 특히 플로우의 첫 번째 request 패킷들을 대상으로 시그니쳐를 적용 함으로써 단독으로 사용할 수 없는, 단순한 트래픽 특징을 조합하여 시그니쳐로 사용할 수 있을 뿐만 아니라 단일 패킷, 단일 플로우에 적용하는 방법보다 정확하게 트래픽을 분석할 수 있다.

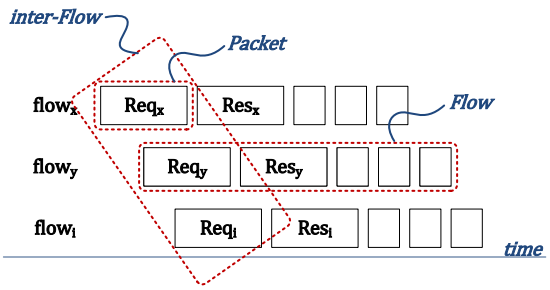

그림 1. 트래픽 분석 단위

행위기반 시그니쳐는 엔트리(Entry)의 조합으로 구성되며 각각의 엔트리는 트래픽의 특징을 가진다. 수식 1, 2 는 각각 행위기반 시그니쳐와 행위기반 시그니쳐를 구성하는 엔트리를 나타낸다.

$$BS = \left\{ \begin{array}{c} A, T, I, E_1, E_2, E_3, \ldots, E_n | \\ n \geq 2, \\ \mathrm{Src}(E_1) = \mathrm{Src}(E_2) = \cdots = \mathrm{Src}(E_n) \end{array} \right\} \quad (1)$$

$$E = \{2^F | F = \{ip, port, prot, payload\}, E \neq \emptyset\} \quad (2)$$

행위기반 시그니쳐(BS)는 응용이름(A), 타입(T), 인터벌(I), 2 개 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 IP 와 목적지 포트 번호는 의미가 없기 때문에 엔트리의 원소에서 제외한다(“any”로 표기). 행위 시그니쳐는 특정 호스트를 기준으로 추출, 적용되기 때문에 모든 엔트리의 출발지 IP 는 동일하여야 한다.

표 1 은 행위기반 시그니쳐의 각 속성에 대한 설명을 나타낸다. 응용 이름(A)은 해당 시그니쳐로 분석된 트래픽에 분석 결과를 명명하기 위해 기술된다. 타입(T)은 Seq(Sequence)와 Set 타입이 있다. Seq

표 1. 행위기반 시그니쳐 속성 및 설명

속성		설명
A		해당 시그니쳐로 분석된 트래픽의 응용 이름
T		엔트리 적용 방법 Seq(Sequence), Set(Set)
I		모든 엔트리가 트래픽에 적용되는 기간(ms)
E	ip	CIDR 표기법의 목적지 IP
	port	목적지 포트 번호
	prot	전송계층 프로토콜 TCP, UDP
	payload	첫 N 바이트 페이로드 HTTP : 10 bytes 이상, Non-HTTP : 2 Byte 이상
Src(Ex)		엔트리 x 의 출발지 IP

는 엔트리들의 순서와 복수 플로우에서 추출한 엔트리가 정확하게 일치되는 것을 의미하고 Set 은 순서에 상관 없이 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 IP 와 포트 번호는는 해당 엔트리가 전송되는 목적지 IP 주소와 포트 번호를 의미하며, IP 의 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. HTTP 를 사용하는 트래픽의 경우 트래픽의 첫 부분에 위치하는 Method(GET, POST, PUT 등)를 구별하기 위해 페이로드의 첫 10 바이트 이상을 사용하고 Non-HTTP 인 경우, 페이로드의 첫 2 바이트 이상을 사용한다.

4. 추출 알고리즘

본 장에서는 행위기반 시그니쳐 추출 알고리즘을 첫 request 패킷 추출 모듈, 후보 시그니쳐 추출 모듈, 그리고 시그니쳐 선택 모듈로 구분하여 각각의 모듈에 대한 알고리즘을 기술한다.

그림 2 는 각 세부 모듈과 입출력 데이터를 보여준다. 최초, 입력 받은 트래픽에서 첫 request 패킷에서 엔트리를 추출하여 리스트 형태로 구성하고, 해당 리스트에서 모든 엔트리 조합을 후보 시그니쳐로 추출한다. 추출된 후보 시그니쳐 중에서 2 대 이상의 호스트에서 공통으로 발생된 후보 시그니쳐를 행위기반 시그니쳐를 추출한다.

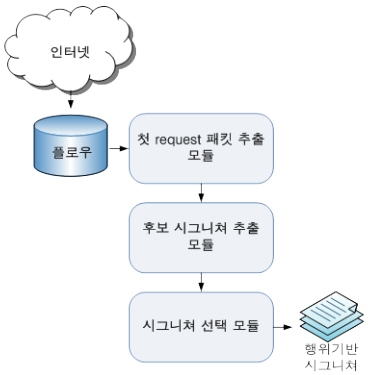

그림 2. 행위기반 시그니쳐 추출 알고리즘

첫 request 패킷 추출 모듈은 플로우 단위로 구분된 패킷들을 입력 받아 각 플로우의 첫 request 패킷에서 행위 시그니쳐 모델에서 정의한 엔트리를 추출하여 리스트로 구성한다. 입력 받은 모든 플로우에서 첫 request 패킷을 통해 엔트리 리스트를 구성하고 시간의 순서로 정렬한다.

후보 시그니쳐 추출 모듈은 앞서 첫 request 패킷 추출 모듈의 출력인 엔트리 리스트를 입력 받아 추출 가능한 모든 후보 패턴을 생성한다. 모든 후보 시그니쳐를 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 시그니쳐를 추출한다. 즉, 입력 받은 엔트리 리스트의 첫 번째 엔트리를 시작으로 최대 인터벌 크기만큼 구간을 설정하고 해당 구간의 엔트리들을 대상으로 최대 엔트리 개수 이내의 추출 가능한 모든 후보 시그니쳐를 추출한다.

시그니쳐 선택 모듈은 앞서 추출된 후보 패턴 중에서 최소 호스트 개수(MIN_PEER)을 초과한 패턴들에 한해 시그니쳐로 선택한다. 행위 기반 시그니쳐는 특정 호스트에 종속되지 않고 모든 호스트에서 특정 응용을 사용할 때 공통으로 발생하는 패턴을 의미한다.

5. 실험 및 평가

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

국내외에서 많은 사람들이 사용하는 응용 5 종(Nateon: 메신져, DropBox: 웹저장소, UTorrent: P2P 파일 전송, Skype: 메신져, Teamviewer: 원격제어)을 선정하였다. 정확한 성능 평가를 위해 4 대의 서로 다른 호스트에서 다른 날짜에 2 회에 걸쳐 다양한 기능을 사용하면서 응용 트래픽을 수집하였다.

표 2. 행위기반 시그니쳐 추출 결과

응용	개수	예시
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/ticker H"), (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 3. 정확도 측정 결과

응용	구분	Precision	Recall
Nateon	flow	1.00 (447/447)	0.60 (447/741)
	byte(K)	1.00 (5064/5064)	0.02 (5064/254110)
DropBox	flow	1.00 (193/193)	0.78 (193/247)
	byte(K)	1.00 (5303/5303)	0.15 (5303/35708)
UTorrent	flow	1.00 (2999/2999)	0.17 (2999/18106)
	byte(K)	1.00 (2741745/2741745)	0.66 (2741745/4182441)
Skype	flow	1.00 (127/127)	0.06 (127/2088)
	byte(K)	1.00 (1589/1589)	0.02 (1589/103342)
Teamviewer	flow	1.00 (239/239)	0.63 (239/385)
	byte(K)	1.00 (8237/8237)	0.04 (8237/215845)
Total	flow	1.00 (4005/4005)	0.18 (4005/21487)
	byte(K)	1.00 (2761938/2761938)	0.57 (2761938/4791446)

5.1 시그니쳐 추출

본 논문에서 제안하는 알고리즘을 통해 추출된 시그니쳐는 표 2 와 같다. 본 실험에서 사용한 임계값은 MAX_INTERVAL 5000ms, MAX_SIZE 10, MIN_PEER 는 4 로 설정하였다.

Nateon 의 경우, 총 48 개의 시그니쳐가 추출되었다. 로그인 시 다른 응용에 비해 다양한 서버(인증 서버, 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 시그니쳐로 추출되었다. 표 2 에 기술된 Nateon 시그니쳐 예시는 총 10 개의 엔트리들로 구성되어있다. Nateon 응용은 서버-클라이언트 형태로 동작하고 고정 포트 번호를 사용하기 때문에 모든 속성을 사용하였다. 제시한 시그니쳐 예시는 특정 인터벌(4324ms)이내에 10 개의 엔트리를 각각 첫 request 패킷으로 가지는 플로우들이 순차적(Seq)으로 발생하는 경우 해당 모든 플로우를 Nateon 으로 분석한다.

UTorrent 의 경우, 총 7 개의 시그니쳐가 추출되었으며, P2P 형태와 임의 포트 번호를 사용하기 때 문에 목적지 IP 와 목적지 포트 번호를 "any"로 표기했다. 제시한 시그니쳐 예시는 특정 인터벌(5000ms) 이내에 2 개의 엔트리가 각각 첫 request 패킷으로 가지는 플로우에 임의(Set)의 순서로 발생하는 경우 모든 해당 모든 플로우를 UTorrent 로 분석한다.

표 4. 페이로드기반 시그니쳐 추출 결과

응용	개수	예시
Nateon	42	.*naeon\.nate\.nate\.com.* ^PVER.*
DropBox	3	^GET /subscribe.host_int=.* .*Dropbox,Inc.*dropbox\.com.*
UTorrent	13	/*BitTorrent protocol.* .*d1:ad2:id20.*
Skype	1	.*User-Agent:.*Skype.*
Teamviewer	1	^..\x00\x17\x24\x6A.\x00.*

5.2 정확도 측정

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

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

TP(True Positive)는 특정 응용 X 의 시그니쳐가 해당 응용 X 를 정확하게 분석된 양을 의미한다. FP(False Positive)는 X 의 시그니쳐가 X 가 아닌 응용을 X 라 분석한 양을 의미하고, FN(False Negative)는 X 의 시그니쳐가 X 를 X 가 아니라고 분석한 양을 의미한다. 즉, Precision 은 해당 응용으로 분석된 트래픽 중에 정확하게 분석된 비율을 의미하고, Recall 은 해당 응용의 전체 트래픽 중에 정확하게 분석된 비율을 의미한다.

표 3 은 응용 5 종의 행위기반 시그니쳐 정확도(precision, recall)를 보여준다. 추출된 모든 시그니쳐는 정확하게 해당 응용을 분석하였다. 즉, 모든 응용 별 Precision 의 값은 1.00 이었다. Recall 의 경우, 플로우 단위 평균 0.18, 바이트 단위 평균 0.57 로 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 시그니쳐로 사용하기 때문에 낮은 Recall 값을 가지지만 시그니쳐의

표 5. 행위기반과 페이로드기반 시그니쳐 분석 결과 비교

응용	구분	Completeness					
		PS	BS	PS∪BS	PS∩BS	PS^c∩BS	PS∩BS^c
Nateon	flow	0.73 (543/741)	0.60 (447/741)	0.73 (543/741)	0.60 (447/741)	0.00 (0/741)	0.13 (96/741)
	byte(K)	0.93 (235,143/254,110)	0.02 (5,064/254,110)	0.93 (235,143/254,110)	0.02 (5,064/254,110)	0.00 (0/254,110)	0.91 (230,079/254,110)
DropBox	flow	0.26 (64/247)	0.78 (193/247)	0.78 (193/247)	0.26 (64/247)	0.52 (129/247)	0.00 (0/247)
	byte(K)	0.01 (68/35,708)	0.15 (5,303/35,708)	0.15 (5,303/35,708)	0.01 (68/35,708)	0.15 (5,234/35,708)	0.00 (0/35,708)
UTorrent	flow	0.79 (14,358/18,106)	0.17 (2,999/18,106)	0.80 (14,488/18,106)	0.15 (2,869/18,106)	0.01 (140/18,106)	0.63 (11,489/18,106)
	byte(K)	0.96 (4,020,339/4,182,441)	0.66 (2,741,745/4,182,441)	0.99 (4,171,534/4,182,441)	0.62 (2,578,702/4,182,441)	0.04 (163,043/4,182,441)	0.34 (1,429,789/4,182,441)
Skype	flow	0.02 (44/2,088)	0.06 (127/2,088)	0.06 (127/2,088)	0.02 (44/2,088)	0.04 (83/2,088)	0.00 (0/2,088)
	byte(K)	0.01 (51/103,342)	0.02 (1,589/103,342)	0.02 (1,589/103,342)	0.01 (51/103,342)	0.01 (1,538/103,342)	0.00 (0/103,342)
Teamviewer	flow	0.01 (1/385)	0.62 (239/385)	0.62 (240/385)	0.00 (0/385)	0.62 (239/385)	0.01 (1/385)
	byte(K)	0.01 (1/215,845)	0.04 (8,237/215,845)	0.04 (8,239/215,845)	0.00 (0/215,845)	0.04 (8,237/215,845)	0.01 (1/215,845)

정확도는 매우 높았다. 따라서, 응용 트래픽 분석(monitoring)측면 보다는 응용 트래픽 탐지 및 제어(detection and control)측면에서 활용이 가능하다. 행위 시그니쳐가 분석한 트래픽의 통계적 특성은 향후 추가적인 연구가 필요하다.

5.3 페이로드기반 시그니쳐와 비교

추출된 행위기반 시그니쳐의 성능을 평가하기 위해 페이로드 시그니쳐로 분석된 결과와 비교 실험을 수행하였다. 비교를 위해 사용한 페이로드 시그니쳐는 LCS 알고리즘을 이용하여 생성하였다. 생성된 페이로드기반 시그니쳐는 표 4 와 같다.

표 5 는 행위기반 시그니쳐와 페이로드 시그니쳐를 동일한 트래픽에 적용한 후, 분석된 결과를 나타낸다. 성능을 확인하기 위한 수식은 다음과 같다.

$$Completeness = \frac{Identified\ Traffic}{Total\ Traffic} \quad (5)$$

“PS”는 페이로드 시그니쳐로 분석한 비율을 나타내고 “BS”는 행위기반 시그니쳐로 분석한 비율을 나타낸다. “PS∪BS”는 두 시그니쳐로 분석된 총 비율을 나타내며 “PS∩BS”는 두 시그니쳐가 공통적으로 분석한 비율을 나타낸다. 또한, “PS^c∩BS”는 페이로드 시그니쳐가 분석하지 못한 트래픽을 행위기반 시그니쳐가 분석한 비율이고, “PS∩BS^c”는 그 반대의 경우이다.

Nateon 의 “PS^c∩BS” 값은 0 이다. 이는 프로토콜 암호화를 거의 사용하지 않는 Nateon 의 특성 때문에 페이로드 시그니쳐가 분석하는 트래픽이 행위기반 시그니쳐가 분석하는 트래픽을 모두 포함하기 때문이다. 이와 대조적으로 Dropbox, Teamviewer, skype 는 행위기반 시그니쳐가 페이로드 시그니쳐를 포함한다. Dropbox 는 데이터 암호화를 위해 HTTPS 트래픽을 많이 발생하는 특성이 있다. LCS 를 통해 HTTPS 트래픽이 시그니쳐로 생성되더라도 이를 단독으로 사용하기에는 정확도가 떨어진다. 하지만, 행위기반 시그니쳐는 여러 엔트리의 조합으로 사용하기 때문에 매우 정확하게 트래픽을 분석한다.

6. 결론 및 향후 과제

본 논문에서는 복수 플로우의 첫 request 패킷에서 트래픽 특징을 추출하여 행위기반 시그니쳐를 추출하는 방법을 제시하였다. 이는 기존 패킷 단위 및 플로우 단위 트래픽 분석의 한계점을 보완한다. 제안한 행위기반 시그니쳐의 타당성을 증명하기 위해 국내외 응용 5 종의 시그니쳐를 추출하고 정확도를 추출하였다. 비록 Recall 측면에서는 낮은 값을 보이는 응용이 존재했지만, 모든 응용에서 100% Precision 을 보였다. 이는 분석된 트래픽은 정확하게 분석되었다는 의미이다. 또한, 기존 LSC 기반 페이로드 시그니쳐와 비교 실험을 통해 행위기반 시그니쳐의 성능을 타당성을 증명하였다.

향후 연구로써는 추출된 시그니쳐가 응용의 어떤 기능을 탐지하는지 확인하는 "Function Naming" 모듈을 추가하는 연구를 진행할 계획이다.

7. 참고 문헌

[1] Myung-Sup Kim, Young J.Won, James Won-Ki Hong, "Application-Level Traffic Monitoring and an Analysis on IP Networks," ETRI Journal Vol.27, No.1, February, 2005.

[2] S. Sen, J. Wang, "Analyzing peer-to-peer traffic across large networks," Internet Measurement Conference (IMC), Proc. Of the 2nd ACM SIGCOMM Workshop on Internet measurement, pp.137-150, 2002.

[3] Karagiannis, T., Papagiannaki, K., Faloutsos, M., "BLINC: Multilevel Traffic Classification in the Dark," ACM SIGCOMM 2005, August/September 2005.

[4] L. Bernaille, R. Teixeira, and K. Salamatian. Early Application Identification. In The 2nd ADETTI/ISCTE CoNEXT Conference, Lisboa, Portugal, December 2006.