멀티 레벨 기반의 응용 트래픽 분석 방법

준회원 오 영 석*, 박 준 상*, 윤 성 호*, 박 진 완*, 이 상 우*, 정회원 김 명 섭*°

Multi-Level basd Application Traffic Classification Method

Young-suk Oh*, Jun-sang Park*, Sung-ho Yoon*, Jin-wan Park*
Sang-woo Lee* *Associate Members*, Myung-sup Kim*° *Regular Member*

요 약

최근 네트워크의 고속화와 인터넷 사용자의 증가에 따른 네트워크 망의 트래픽 급증으로 네트워크 자원의 효율적인 관리와 응용 기반 트래픽 분석의 중요성이 갈수록 강조되고 있다. 이미 기존의 많은 논문들에서 효율적인 네트워크 자원 관리를 위한 응용 프로그램 별 트래픽 분석에 대한 다양한 방법론과 알고리즘을 제안하고 있지만 각각의 연구는 한계점을 가지고 있다. 본 논문에서는 멀티 레벨 기반의 응용 트래픽 분석 방법론을 제안한다. 본 연구는 Header, Statistic, Payload 시그니쳐 기반 개별 분석 방법론과 Behavior 알고리즘을 이용한 방법론의 결과를 바탕으로 트래픽 상관관계를 적용하여 추가적인 분석이 가능하게 한다. 각각의 분석 방법론을 통합하여 기존 하나의 분석 시스템이 가지는 단점을 보완함으로써 유연하고 견고한 멀티 레벨 분석 시스템을 구축하였다. 또한 검증 시스템을 통해 학내 네트워크에 적용하여 그 타당성을 증명하였다.

Key Words : Multi-Level Traffic Classification, Signature-based, Correlation Algorithm

ABSTRACT

Recently as the number of users and application traffic is increasing on high speed network, the importance of application traffic classification is growing more and more for efficient network resource management. Although a number of methods and algorithms for traffic classification have been introduced, they have some limitations in terms of accuracy and completeness. In this paper we propose an application traffic classification based multi-level architecture which integrates several signature-based methods and behavior algorithm, and analyzes traffic using correlation among traffic flows. By strengthening the strength and making up for the weakness of individual methods we could construct a flexible and robust multi-level classification system. Also, by experiments with our campus network traffic we proved the performance and validity of the proposed mechanism.

I. 서 론

최근 네트워크의 고속화와 더불어 전화망이나 전용망 기반의 음성 및 영상 서비스가 패킷 기반의 IP Network에 통합되고, 다양한 서비스와 응용프로그램이 개발됨에 따라 기업이나 개인들은 인터넷으로 대표되는 네트워크에 대한 의존이 상당히 커져가고 있고 인터넷을 기반으로 하는 On-line 산업의 범위도 증가하고 있다. 이와 같은 현실 속에서 네트워크의 효율적 운용과 관리를 위한 트래픽의 모니터링과 분석은 네트워크 사용 현황 파악과 확장 계획 수립 등의 전통적인 필요성 외에 다양한 분야에서 커져가고 있다.

※ 이 논문은 2007년 정부(교육인적자원부)의 재원으로 한국학술진흥재단(KRF-2007-331-D00387)과 2009년 정부(교육과학기술부)의 재원으로 한국연구재단(2009-0090455)의 지원을 받아 수행된 연구임.

* 고려대학교 컴퓨터정보학과 ({youngsuk_oh, junsang_park, sungho_yoon, jinwan_park, sangwoo_lee, tmskim}@korea.ac.kr), (° : 교신저자)

논문번호 : KICS2010-06-275, 접수일자 : 2010년 6월 21일, 최종논문접수일자 : 2010년 7월 19일

네트워크 트래픽 분석에 있어 해결되어야 할 많은 문제들이 있다. 그 중에도 가장 중요하고 선행되어야 할 문제는 다양한 응용 트래픽들을 어떻게 분류할 것인가의 문제와 트래픽 분석을 위해 제시된 여러 방법론들을 검증 할 수 있는 표준화된 시스템을 구축하는 것이다. 이 부분의 해결은 그 후에 이어지는 다양한 분석들의 결과에 대한 신뢰성을 결정하고 트래픽과 인터넷 사용자의 상관관계를 이해하는 데 매우 중요하다. 특히 응용 레벨 트래픽의 정확한 분류 방법에 관한 연구는 연계된 활용들(응용 별 종량제 과금, 응용기반 트래픽 제어, CRM, SLA 지원, 응용 레벨 트래픽 보안 등)의 분석 신뢰성을 결정하는 중요한 요소이다[1,2].

본 논문의 최종 목표는 기존 연구인 시그니쳐 기반 분석 방법으로 분류되지 않는 트래픽을 멀티 레벨 분석 방법을 적용하여 보다 신뢰성 높은 분석 방법을 개발하는데 있다. 본 논문에서는 각 분석 방법의 성능 평가 결과를 바탕으로 각 분석 방법의 장점들을 취합하여 보다 신뢰성 높은 분석 방법을 제안한다. 본 논문에서 제안하는 방법론은 세 가지의 시그니쳐 기반 분석 방법과 두 가지의 알고리즘 기반 방법으로 구성된다. 우선 Header 시그니쳐[12]와 Statistic 시그니쳐[13], Payload 시그니쳐[14] 기반 분석과 추가적으로 시그니쳐에 기반한 분석 방법에 적용이 불가능한 응용에 대하여 개별 응용 프로그램의 트래픽 발생 형태에 기반한 분석 방법을 적용한다. 위 방법을 이용하여 1차적으로 응용 트래픽을 분류하고, 분류되지 않은 응용 트래픽에 대하여 트래픽의 상관관계 기반한 분류 방법을 적용하는 멀티 레벨 분석 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 트래픽 분석에 관한 기존 연구들을 살펴보고 3장에서는 본 논문에서 제안하는 멀티 레벨 분석 시스템에 포함되는 세 가지의 시그니쳐 추출 및 시그니쳐 기반 분석 알고리즘과 Behavior 기반 알고리즘 그리고 Correlation 기반 분석 알고리즘의 전반적인 내용을 설명한다. 4장에서는 멀티 레벨 분석 시스템의 전체적인 구성과 분석 과정에 대하여 기술한다. 5장에서는 앞서 제안한 분석 시스템으로 분석한 결과에 대한 검증 방법을 기술하고 학내 망에 적용한 결과를 제시한다. 마지막으로 6장에서는 결론 및 향후 과제에 대해 기술한다.

Ⅱ. 관련 연구

이미 많은 기존 연구에서 응용 트래픽을 분석하기 위해 다양한 방법론들을 제안되어왔다. 각각의 분석 방법론들은 그에 따른 한계점을 가지고 있고 계속적으로 다양한 응용들이 출현하고 변하기 때문에 응용 별 트래픽 분석에 대한 꾸준한 연구가 진행되고 있다. 트래픽 분석 방법은 크게 포트 기반[3,4,5,6,16], 시그니쳐[7,14], 머신 러닝[8,9,16], 트래픽 상관관계[1,4,10]기반 분석으로 나눌 수 있다.

2.1 포트 기반 트래픽 분석

과거의 인터넷 트래픽 분석은 잘 알려진 포트 번호(well-known Port Number)를 사용하는 HTTP, FTP, e-mail, SMTP 등 IANA[5]에서 지정한 포트 정보를 이용하였다. 하지만 최근 사용되는 응용들은 방화벽 및 IPS장비를 통과하기 위해 잘 알려진 포트 번호(well-known Port Number)의 사용을 피하고 있다. 따라서 포트 기반 분석은 더 이상 높은 신뢰성과 분석률을 제공할 수 없게 되었다.

2.2 시그니쳐 기반 트래픽 분석

시그니쳐 기반의 방법[7,14]은 각 응용 트래픽 별로 그들만이 사용하는 다른 응용들과 구분되는 공통분모를 찾아내어 그것을 이용해 트래픽을 분류하는 방법이다. 이 방법은 시그니쳐가 확인된 응용에 대해서는 정확한 분석이 가능한 장점을 갖는다. 하지만 모든 응용별로 수작업을 통해 시그니쳐를 찾아야 하고 시그니쳐를 확인하기 힘든 응용들이 존재하며, 찾아진 시그니쳐가 응용의 변화에 적절히 대처하지 못한다는 단점이 있다.

2.3 머신 러닝 기반 트래픽 분석

머신 러닝 기반 분석 방법[8,9,16]은 응용 별 트래픽의 특징이 될 수 있는 요소(port, inter-arrival time, packet size)를 머신 러닝 알고리즘으로 학습을 시킨 후에 분석하는 방법이다. 이 방법의 장점은 고급 알고리즘을 이용하여 트래픽을 분석하기 때문에 다른 분석 방법에 비해 분석률이 높다는 것이다. 하지만 제한된 범위의 응용에 한하여 트래픽을 수집하고 미리 학습해야 하는 사전 준비 작업의 오버헤드가 있고 실제 네트워크에 적용하였을 경우에도 분석의 정확성이 떨어지는 단점을 가지고 있다.

2.4 트래픽 상관관계 기반 분석

본 방법론은 인터넷 트래픽의 3레벨 주소체계(IP address, port number, protocol)과 트래픽 발생 형태 등의 고유한 특성을 바탕으로 연관성을 가중치로 표현하고 그 임계값을 설정하여 트래픽을 분석하는 방법이다. 트래픽을 분석하는 관점에서 응용들이 가지는 특징을 분석에 활용하기 때문에 높은 분석률을 가지는

장점이 있는 반면, 응용 별 특징의 활용에 대한 명확한 알고리즘 없이 trial-and-error의 방법으로 임계값을 찾는 방법이기 때문에 실제 인터넷 트래픽에 적용하였을 경우 분석 결과에 대한 신뢰성을 보장하기 어렵다.

본 연구에서 제안하는 멀티 레벨 기반 분석 방법은 시그니쳐 기반 분석 방법론의 확장에 해당된다. 본 분석 방법론은 Header, Statistic, Payload 시그니쳐 분석 방법과 Behavior 알고리즘, Correlation 알고리즘을 적절히 조합하여 멀티 레벨 분석 시스템을 구축한다. 각 분석 시스템이 분석하지 못하는 트래픽을 트래픽 간의 상관관계를 이용하여 분석할 수 있다. 본 연구에서는 각 하나의 분석 방법에 대한 단점을 보완하고 실제 학내 네트워크에 적용하여 그 타당성을 증명한다.

Ⅲ. 시그니쳐 추출 및 분석 알고리즘

본 장에서는 먼저 시그니쳐 추출 시스템[11,12,13]과 시그니쳐 분석 알고리즘에 대하여 설명한다. 이어서 실제 트래픽을 분석하는 멀티 레벨 분석 알고리즘에 대해 기술한다.

그림 1은 세 가지 시그니쳐 추출 시스템의 전체적인 구성도를 보여준다. 각 추출 시스템은 Header[11], Statistic[12], Payload[13] 시그니쳐를 생성하여 xml 파일 형식으로 저장한다. 트래픽은 특정 네트워크에서 인터넷으로 향하는 모든 트래픽을 수집한다. 수집한 패킷을 이용하여 FG(Flow Generator)가 플로우를 생성한다. TMS는 네트워크 내의 TMA[15]가 설치된 여러 단말 호스트들로부터 TMA로그를 수집하고 GTG(Ground Truth Generator)에게 전달한다. GTG는 앞서 FG가 생성한 플로우와 TMS로부터 전달받은 TMA 로그 정보를 비교하여 Ground Truth Traffic Data를 생성한다. 이 GT(Ground Truth) 정보를 바탕으로 HSG, SSG, PSG 는 각각의 시그니쳐를 생성한다. 또한 GT를 바탕으로 BAG는 분석 알고리즘을 구축한다. 각 시그니쳐는 각 분석 시스템(HSC, SSC, PSC)의 입력 데이터로 들어가 각각의 분석 결과를 Flow형태로 출력한다. 각 분석 결과는 Correlation기반 분석 시스템의 입력 데이터로 사용되어 최종적인 분석 결과를 추출하게 된다. 각각의 분석 방법에 대한 자세한 내용은 본 장의 3절에서 설명한다.

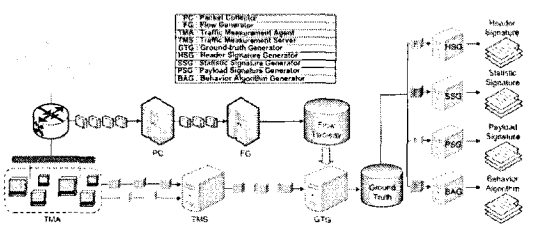


그림 1. 시그니쳐 추출 시스템의 전체 구성도
Fig. 1. Signature Generation System Configuration

3.1 시그니쳐 기반 분석 알고리즘

본 절에서는 시그니쳐 추출 시스템[11-13]에 의해 생성된 Header, Statistic, Payload 시그니쳐를 이용한 트래픽 분석에 대해 설명한다.

그림 2는 세 가지 시그니쳐 기반의 분석 알고리즘의 흐름도를 보여준다. 패킷이 들어오면 해당 패킷의 플로우가 생성되어 있는지 확인하고 업데이트를 한다. 이미 확인된 플로우라면 다음 패킷을 검사하고 결정되지 않은 플로우이면 시그니쳐 분석 시스템의 세 가지 (Header, Statistic, Payload)분석 시스템의 시그니쳐 리스트와 실제 플로우를 비교함으로써 분석한다. 세 가지 시그니쳐 리스트 중 하나라도 매칭이 되면 그 플로우는 분석된 해당 응용의 플로우로 결정되고 세 가지 리스트 모두 매칭이 되지 않으면 어떤 응용의 트래픽인지 결정을 할 수 없으므로 다음 패킷을 검사한다. 매칭 되지 않은 플로우는 Correlation 알고리즘을 이용하여 분석된다. Correlation 알고리즘은 3.3절에서 설명한다.

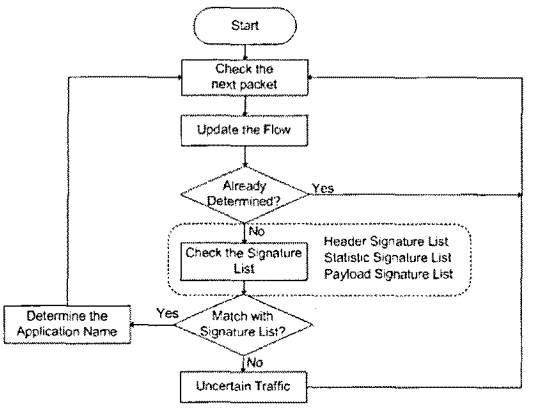


그림 2. 시그니쳐 분석 알고리즘 흐름도
Fig. 2. Signature Classification Algorithm Flow Chart

3.2 Behavior 기반 분석 알고리즘

Skype 응용 트래픽 탐지 알고리즘의 핵심은 트래픽 플로우의 초기 몇 개 패킷의 DPI 분석을 통해 탐지하고 이를 바탕으로 SC(Skype Client)가 설치된 호스트의 {IP, port}리스트를 구축함으로써 다른 SC로부터 발생하는 Skype 플로우를 쉽게 탐지하는 것이다.

그림 3은 Skype 응용 트래픽 탐지 알고리즘의 순서

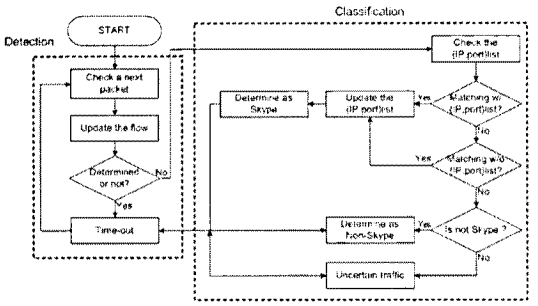


그림 3. Skype 응용 Behavior 기반 탐지 알고리즘
Fig. 3. Skype Detection Algorithm based Behavior

도이다. 먼저 패킷이 캡쳐되면 해당 패킷을 기반으로 플로우 정보를 생성 또는 이미 생성된 플로우의 경우 플로우 정보를 갱신한다. 플로우가 이미 Skype 플로우인지 아닌지 이미 결정된 상태라면 더 이상 검사를 수행하지 않고 다음 패킷을 기다린다. 하지만 아직 결정되지 않은 플로우라면 Classification 단계에 들어서게 된다. Classification

단계에서는 해당 플로우가 Skype 응용인지 아닌지 결정을 내리거나 더 많은 패킷들을 살펴봐야 한다고 (Uncertain traffic)판단을 내리게 된다. 또한 해당 플로우가 Skype 응용의 플로우로 판단 시 새롭게 생성된 호스트의 {IP, port}정보를 리스트에 추가한다.

3.3 Correlation 기반 분석 알고리즘

본 절에서는 멀티 레벨 기반 분석 시스템의 핵심이 되는 Correlation 기반 분석 알고리즘에 대해 설명한다.

3.3.1 정의

트래픽 상관관계 기반 분석 방법은 인터넷 트래픽의 3-tuple레벨 주소 체계(IP address, port number, transport protocol)와 트래픽 발생 시점, 형태 등의 특성을 바탕으로 트래픽 플로우들 사이에 연관성을 가중치로 표현하고 임계값을 적용하여 트래픽을 응용 별로 구분하는 방법이다. 이 방법은 시그니쳐 기반 분석 방법으로 95% 이상의 트래픽을 정확하게 분석 가능하기 때문에 플로우 사이의 다양한 연관성을 찾을 수 있으며 이를 바탕으로 분석률을 향상 시킬 수 있다. 본 논문에서는 다음 세 가지의 트래픽 상관관계를 찾아 실제 트래픽에 적용하였다.

3.3.1.1 서버-클라이언트 기반 상관관계(Server-Client)

본 방법론은 하나의 서버 3-tuple(IP address, port number, transport protocol)을 동시에 하나의 응용만 사용한다라는 가정을 바탕으로 한다. 만약 선행 시그니쳐 기반 분석 방법으로 분석된 트래픽 중 서버를 포함하는 분석 결과를 가지고 있다면, 해당 서버의 3-tuple을 해당 응용이 사용한 것으로 간주하고 분석 대상이 되는 네트워크에 해당 서버와 통신하는 트래픽을 해당 응용으로 분석한다.

3.3.1.2 발생시간 기반 상관관계(Time-based)

본 방법론은 한 호스트에 한해 일정 기간 안에 발생되는 트래픽은 같은 응용일 가능성이 높다는 가정에서부터 시작한다. 선행 방법으로 분석되지 않는 트래픽들을 호스트와 짧은 시간 간격으로 그룹한 후, 해당 그룹 중 분석된 트래픽이 있을 경우 그 그룹에 속한 모든 트래픽을 해당 응용으로 분석하는 방법이다. 이 방법은 그룹을 하는 기준이 일정 기간에 따라 분석률과 정확도가 영향을 받기 때문에 해당 네트워크에 맞는 적절한 기간을 설정하기 위한 실험이 필요하다.

3.3.1.3 호스트-호스트 기반 상관관계(Host-Host)

본 방법론은 두 호스트 간에 일어나는 통신은 하나의 응용에 의해 발생될 가능성이 높다는 가정으로부터 시작한다. 즉, 호스트 간의 통신에서 일 부분만 선행 방법으로 분석이 되었다면 분석되지 않은 트래픽을 분석된 트래픽의 응용으로 분석하는 것이다. 데이터를 주고 받는 통신일 경우 제어를 위한 통신과 실제 데이터를 전송하기 위한 통신이 서로 다른 포트에서 이루어 진다. 따라서 이러한 형태를 보이는 호스트들 사이의 트래픽을 호스트-호스트 상관관계를 이용하여 효과적으로 분석한다.

그림 4는 플로우 상관관계 기반 세 가지 방법의 우선순위를 보여준다. 상관관계 기반 분석 방법에서는 미 분석된 트래픽만을 분석 대상으로 하므로 정확도가 가장 큰 Server-Client기반 상관관계 방법론의 우선순위를 높게 부여하여 시그니쳐 분석 결과의 정확도를 떨어뜨리지 않고 추가적인 분석을 가능하게 한다. 본 알고리즘의 가장 큰 장점은 선행되는 시그니쳐 기반 분석 시스템이 분석하지 못하는 트래픽을 상관관계를 기반으로 추가적으로 분석할 수 있다는 점이다. 플로

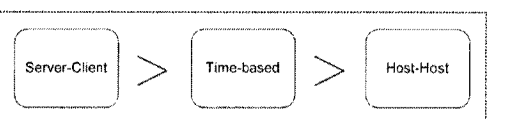


그림 4. 상관관계 기반 방법론의 우선순위
Fig. 4. Correlation Mechanism Priority

우 상관관계 기반 분석 알고리즘은 전제 조건이 있다. 그것은 정확하게 분석된 분석 결과가 있어야 한다는 것이다. 상관관계를 이용하면 서로 연관성이 강한 플로우들끼리 집합으로 묶어지지만, 그 중 하나라도 분석된 결과가 있어야 해당 집합에 속한 플로우 전체를 분석할 수 있는 것이다. 따라서 앞선 분석 결과에 따라 플로우 상관관계 분석을 위해 시그니쳐 기반의 분석 방법이 선행되고 또한, 그 결과 역시 95% 이상의 분석률을 보이고 있어 선행 분석 결과를 보완한다는 측면에서 그 의미를 찾을 수 있다. 시그니쳐 기반 분석 방법은 해당 트래픽에 매칭되는 시그니쳐를 가지고 있지 않으면 분석하지 못하거나 잘못 분석하는 제한이 있다. 하지만 플로우 상관관계를 이용하면 시그니쳐를 가지고 있지 않은 트래픽도 플로우 간의 상관관계를 이용하여 추가적으로 분석이 가능하다. 상관관계 기반 분석 알고리즘을 통해 기존에 찾지 못한 응용들은 대부분 페이로드가 없는 플로우들이다. 즉, 모든 응용은 페이로드가 없는 플로우와 페이로드가 존재하는 플로우 모두를 포함하기 때문에 상관관계 기반 분석 알고리즘으로 인하여 기존에 찾지 못한 응용들은 대부분 페이로드가 없는 플로우들이므로 대부분의 응용에서 상관관계를 통해 분석이 가능한 것이다.

Ⅳ. 멀티 레벨 분석 시스템

본 장에서는 네 가지 트래픽 분석 시스템과 트래픽 상관관계 기반의 분석 알고리즘을 포함하는 멀티 레벨 기반 트래픽 분석 시스템에 대한 전반적인 구성에 대해 설명한다.

4.1 전체 구성도

본 절에서는 멀티 레벨 분석 시스템의 전체적인 구조를 살펴본다.

그림 5는 본 논문에서 제안하는 멀티 레벨 분석 시스템의 전체적인 구성도이다.

멀티 레벨 분석 시스템은 크게 네 단계로 구성되어 있다. 첫 번째 단계(1-level)에서는 라우터에서 수집된 모든 패킷들을 FG(Flow Generation)에 의해 플로우(Flow)형태로 변환된다. 생성된 모든 Flow들은 두 번째 단계(2-level)에서 각 시그니쳐 분석 시스템(Header[11], Statistic[12], Payload[13] Signature Classifier)과 Behavior 알고리즘 분석 시스템의 입력 데이터로 사용된다. 각 분석 시스템(HSC, SSC, PSC, BAC)은 매 분마다 분류되지 않은 플로우를 동일하게 전달받는다. 3.1절에서 설명한 바와 같이 시그니쳐 기반 분석

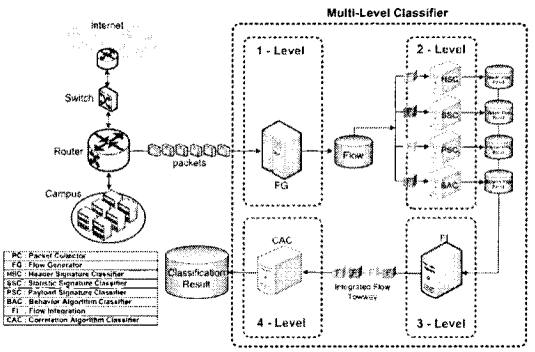


그림 5. 멀티 레벨 트래픽 분석 시스템 구성도
Fig. 5. Multi-Level Traffic Classification System

시스템은 각각의 시그니쳐를 입력으로 하여 실제 플로우를 비교하여 시그니쳐의 매칭을 통해 플로우의 응용을 확인한다. Behavior 기반 분석 시스템 또한 3.2절에서 설명한 바와 같이 Skype 응용 프로그램의 동작형태에 기반 한 알고리즘을 통하여 Skype 응용이 발생시킨 플로우를 분석한다. 해당 응용으로 분석된 플로우나 분석되지 않은 플로우들은 세 번째 단계(3-level)에서 FI(Flow Integration)에 의해 통합되어 진다. 3-Level과 4-Level을 합하여 Correlation System이라고 한다. 크게 'Integration + Correlation'으로 나눌 수 있는데 3-Level에서 실제로는 Header[11], Statistic[12], Payload[13] 각 개별 분석 시스템의 우선순위를 반영한다. 마지막 네 번째 단계(4-level)에서는 CAC(Correlation Algorithm Classifier)가 앞선 단계(3-level)에서 통합된 플로우를 입력으로 하여 두 번째 단계(2-level)에서 분석되지 않은 플로우들을 3.3절의 세 가지 상관관계 알고리즘을 이용하여 추가적인 분석을 수행한다.

Ⅴ. 평가 및 검증

본 장에서는 멀티 레벨 기반 분석 시스템의 타당성을 증명하기 위해 객관적인 검증을 위한 평가 요소를 기술하고 학내 네트워크의 실제 트래픽에 적용한 결과에 대해서 설명한다.

5.1 평가 요소

본 연구에서 개발한 멀티 레벨 분석 시스템의 결과를 검증하기 위해 다음과 같은 평가 요소를 정의하고 사용하였다. 검증 평가 요소는 범위(Coverage), 분석률(Completeness), 정확도(Accuracy)로 구성되고 각 평가 요소는 다각적인 분석을 위해 Flow, Packet, Byte 단위로 각각 표시한다.

첫 번째 검증 평가 요소인 범위(Coverage)는 해당 분석 시스템이 분석 가능한 응용의 개수를 나타낸다. 두 번째 요소인 분석률(Completeness)은 해당 분석 시스템이 전체 트래픽 중에 분석한 결과의 양을 비율로서 나타내는 것이다. 마지막 평가 요소인 정확도(Accuracy)는 해당 분석 시스템이 분석한 결과를 Ground Truth로 검증하여 얼마나 정확하게 분석하였는지에 대해 비율로서 나타낸다.

5.2 검증 결과

본 논문에서 제안한 멀티 레벨 기반 분석을 위해학내 네트워크 트래픽을 대상으로 실험을 하였다. 본 실험은 Intel Dual-Core E2140 1.60GHz CPU와 3GB RAM이 탑재된 범용 컴퓨터에서 5일간 학내 망에서 발생되는 연속적인 전체 트래픽을 바탕으로 분석 결과를 도출하였다.

본 논문에서 제안한 멀티 레벨 분석 시스템의 성능을 평가하기 위해 표 1에 기술한 트래픽 트레이스를 바탕으로 실험을 하였다. 본 실험은 Day 1-4동안 세 가지(Header, Statistic, Payload) 시그니쳐와 Skype 응용에 대한 Behavior알고리즘을 추출한 후 Day5의 트래픽에 대해 추출과 분석을 동시에 수행하였다. 본 논문에서 사용된 시그니쳐들은 본 연구에서 직접 개발한 자동 추출 시스템[11-13]에 의해 생성되었다. 또한 추출된 시그니쳐들은 정확성을 향상시키고 유효성을 높이기 위해 수작업을 통하여 수정되었다. 즉, 기존 개별 분석 시스템과 멀티 레벨 분석 시스템에 사용된 시그니쳐들은 동일하다.

표 1. 실험에 사용된 트래픽 트레이스
Table 1. Traffic Trace

	Flow(K)	Packet(M)	Byte(G)
Day1	3,089/48,872	111/1,801	86/1,541
Day2	3,220/21,767	95/749	67/624
Day3	3,352/55,940	90/2,034	65/1,707
Day4	3,450/53,353	136/1,969	112/1,673
Day5	2,932/52,282	95/6,051	70/58,334

5.2.1 Coverage

표 2는 본 분석 시스템의 Coverage를 나타내고 그림 6은 Coverage의 세부적인 내용을 보여 준다.

본 분석 시스템의 Coverage 즉, 본 분석 시스템으로 분석한 응용의 개수를 살펴보면 총 응용의 개수는 246개이고 총 프로세스 기준으로 보면 384개로 알 수 있다.

그림 7은 본 멀티 레벨 분석 시스템을 통해 분석을 수행한 총 응용의 개수를 Header, Statistic, Payload

표 2. Coverage
Table 2. Coverage

	Coverage
Process	246
Signature	384

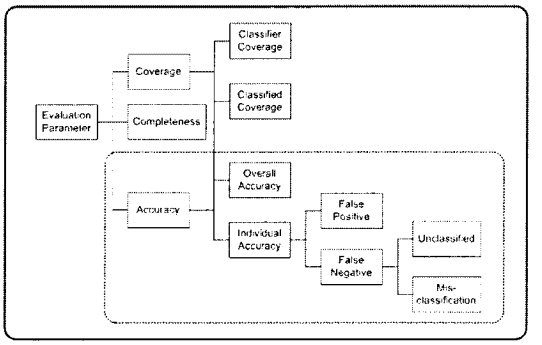


그림 6. 검증 평가 요소
Fig. 6. Verification Evaluation Elements

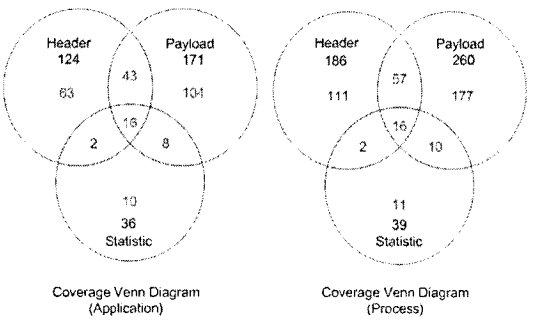


그림 7. 분석 응용 범위
Fig. 7. Application Coverage of Classification

시그니쳐 별로 벤 다이어그램으로 나타낸 것이다. Payload, Header, Statistic 시그니쳐 별로 Coverage가 많이 차지하는 것을 알 수 있다.

5.2.2 Completeness

표 3은 본 논문에서 제안하는 멀티 레벨 기반 분석 시스템으로 분석한 분석률과 Header[11], Statistic[12], Payload[13] 시그니쳐 기반의 네 가지 분석 시스템의 분석률을 보여준다.

멀티 레벨 기반의 분석 결과를 다른 시그니쳐 기반의 분석 결과와 비교해 보았을 때, 본 방법론의 분석률이 상대적으로 높은 것을 알 수 있다. 분석률이 낮은 Header, Statistic 시그니쳐 기반 분석 방법론의 단점을 본 방법론의 Correlation 기반 분석 시스템에 의해 해결할 수 있었다. 즉, Correlation 기반 분석 시스템은

표 3. Multi-Level 분석 시스템의 분석률
Table 3. Completeness of Multi-Level Classification System

	Week Completeness		
	Flow	Packet	Byte
Multi-Level	97.63%	95.47%	95.69%
Header Signature	24.83%	5.15%	3.99%
Statistic Signature	18.99%	30.31%	31.52%
Payload Signature	86.06%	76.75%	75.72%

네 가지의 방법론 각각의 분석 결과를 이용하여 상관관계 기반의 알고리즘을 이용하기 때문이다. Header, Statistic, Payload 시그니쳐 기반 분석 방법과 Behavior 알고리즘으로 분석하지 못한 트래픽을 플로우 간의 상관관계를 기반으로 분석하므로 분석률은 향상된다.

5.2.3 Accuracy

표 4는 본 멀티 레벨 분석 시스템의 정확도와 기존의 선행 분석 시스템(Header[11], Statistic[12], Payload[13] Signature Classifier) 방법론의 정확도를 보여 준다.

위 결과와 같이 본 분석 시스템은 다른 기존의 분석 시스템과 같이 97% 이상의 높은 정확도를 보여준다. 즉, 본 시스템은 개별 분석 시스템의 높은 정확도를 유지하고 있다. 하지만 본 분석 시스템의 정확도는 Header 시그니쳐 기반 분석 결과에 비해 약간 떨어지는 것을 알 수 있다. 그 원인은 대부분 다른 응용 간의 충돌로 인해 발생한다. 충돌이 발생한 해당 트래픽을 분석한 결과 하나의 응용이 여러 응용 레벨 프로토콜을 사용한 경우가 많았다. 특히 같은 응용 레벨 프로토콜을 사용하는 Web Disk 응용 프로그램들 간의 충돌이 정확도를 떨어뜨리는 주된 원인 이었다. 또한 다양한 응용프로그램들이 Web기반 서비스를 제공하기 때문에 Web 기반 응용과 Internet explore와의 충돌도 오 분류의 원인이 되었다.

표 4. Multi-Level 분석 시스템의 정확도
Table 4. Accuracy of Multi-Level Classification System

	Week Accuracy		
	Flow	Packet	Byte
Multi-Level	97.94%	97.15%	96.88%
Header Signature	99.91%	99.92%	99.91%
Statistic Signature	97.20%	91.89%	98.11%
Payload Signature	97.49%	94.18%	95.40%

VI. 결론 및 향후 과제

트래픽 분석의 중요성이 네트워크 관리에 있어 점점 강조되고 있다. 기존의 많은 연구를 통해 다양한 트래픽 분석 방법론이 제시되었지만 각각의 방법론들은 한계점을 가지고 있어 실제 네트워크 환경에 적용하는 데 어려움이 있다. 본 논문에서는 Header, Statistic, Payload 시그니쳐 기반의 분석 시스템들[11,12,13]이 분석하지 못하는 트래픽에 대해 실제 플로우 트래픽의 상관관계를 기반으로 한 분석 알고리즘을 개발하였고 이를 바탕으로 선행 분석 시스템으로 분석하지 못하는 트래픽을 대상으로 분석률을 높일 수 있는 견고하고 신뢰성 높은 멀티 레벨 분석 시스템을 구축하였다. 본 시스템은 선행 분석 시스템들의 네 가지 분석 결과를 통합하고 상관관계를 이용한 알고리즘을 통해 각 분석기의 성능을 평가할 수 있었고 각 시그니쳐 기반의 분석 시스템이 가지고 있는 한계점을 보완할 수 있었다. 또한 각 분석 알고리즘의 이점을 적절하게 조합하여 다양한 방법론을 사용하게 되므로 하나의 방법론으로 트래픽을 분석하는 방법보다 정확하고 유연한 분석이 가능하다. 본 멀티 레벨 분석 시스템을 실제 네트워크에 적용한 결과 총 246개의 응용 프로그램에서 발생되는 트래픽을 분석하였고 Flow기준 97.63%의 분석률과 97.94% 정확도를 확인할 수 있었다.

본 논문의 검증 결과 정확도를 떨어뜨리는 응용 프로그램 간의 충돌을 확인할 수 있었다. 따라서 응용 프로그램 간의 충돌 문제에 대해 우선순위를 정하여 정확성을 높일 수 있는 방법에 대한 심층적인 연구가 필요하다. 앞으로 본 멀티 레벨 분석 시스템을 바탕으로 개별 시그니쳐 기반의 각각 분석 시스템의 성능 향상에 관한 연구와 트래픽의 특징 및 상관관계에 대한 연구를 통해 정확성을 높임으로써 견고하고 신뢰성 높은 시스템을 구축할 계획이다.

참 고 문 헌

[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, Feb. 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] W. Li et al. "Efficient application identification and the temporal and spatial stability of classification schema", *Computer Networks,*

2009.doi:10.1016/j.comnet.2008.11.016.

[4] Thomas Karagiannis, Konstantina Papagiannaki, Michalis Faloutsos. "BLINC: Multilevel Traffic Classification in the Dark", *Proc. of SIGCOMM* 2005, Philadelphia, PA, Aug. 22-26, 2005.

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

[6] Jian Zhang and Andrew Moore, "Traffic Trace Artifacts due to Monitoring Via Port Mirroring," *Proc. of the IEEE/IFIP Workshop onEnd-to-End Monitoring Techniques and Services (E2EMON) 2007*, Munich, Germany, May 21, 2007.

[7] Risso, F. Baldi, M. Morandi, O. Baldini, A. Monclus, P. "Lightweight, Payload-Based Traffic Classification: An Experimental Evaluation," *Proc. of the Communications, 2008. ICC '08. IEEE International Conference*, 2008.

[8] Jeffrey Erman, Martin Arlitt, Anirban Mahanti, "Traffic Classification Using Clustering Algorithms," *Proc. of SIGCOMM Workshop on Mining network data*, Pisa, Italy, Sep. 2006, pp.281-286.

[9] Andrew W. Moore and Denis Zuev, "Internet Traffic Classification Using Bayesian Analysis Techniques," *Proc. of the ACM SIGMETRICS*, Banff, Canada, Jun. 2005.

[10] Thomas Karagiannis, Konstantina Papagiannaki, and Michalis Faloutsos. "BLINC: Multilevel Traffic Classification in the Dark," *Proc. of SIGCOMM 2005*, Philadelphia, PA, Aug. 22-26, 2005.

[11] Sung-Ho Yoon, Jun-Sang Park, Jin-Wan Park, Sang-woo Lee, Myung-Sup Kim, "Fixed IP-port based Application-Level Internet Traffic Classification", *정보처리학회논문지* C 제17-C권 제2호, April. 2010, pp.205-214

[12] Jin-wan Park, Sung-ho Yoon, Jun-sang Park, Sang-woo Lee, Myung-sup Kim, "Statistic Signature based Application Traffic Classification", *한국통신학회논문지*, Vol.34, No.11, Nov. 2009, pp.1234-1244.

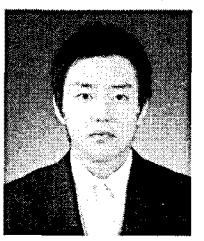
[13] Jun-Sang Park, Jin-Wan Park, Sung-Ho Yoon, Hyun-Shin Lee, Myung-Sup Kim, "Developmen of Signature Generation and Update System for Application-level Traffic Classification" *정보처리학회논문지* C 제17-C권 제1호, Feb. 2010, pp. 99-108.

[14] Liu, Hui Feng, Wenfeng Huang, Yongfeng Li, Xing "Accurate Traffic Classification", *Networking, Architecture, and Storage, 2007.* NAS 2007. International Conference.

[15] Byung-Chul Park, Young J. Won, Myung-Sup Kim, James W. Hong, "Towards Automated Application Signature Generation for Traffic Identification," *Proc. of the IEEE/IFIP Network Operations and Management Symposium (NOMS) 2008*, Salvador, Bahia, Brazil, pp.160-167, April. 7-11, 2008.

[16] Hyun-chul Kim, kc claffy, Marina Fomenkov, Dhiman Barman, Michalis Faloutsos, Ki-young Lee, "Internet Traffic Classification Demystified: Myths, Caveats, and the Best Practices" *Proc. of ACM SIGCOMM CoNEXT*, Madrid, Spain, Dec, 2008

오 영 석 (Young-suk Oh) 준회원

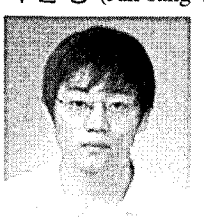


2009년 고려대학교 컴퓨터정보학과 학사

2009년~현재 고려대학교 컴퓨터 정보학과 석사과정

<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

박 준 상 (Jun-sang Park) 준회원



2008년 고려대학교 컴퓨터정보학과 학사

2008년~현재 고려대학교 컴퓨터정보학과 석사과정

<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

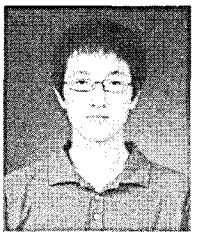
윤 성 호 (Sung-ho Yoon) 준회원

2009년 고려대학교 컴퓨터정보학과 학사

2009년~현재 고려대학교 컴퓨터정보학과 석사과정

<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

박 진 완 (Jin-wan Park) 준회원

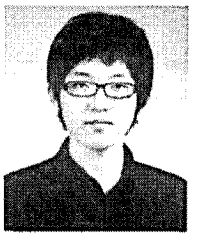


2009년 고려대학교 컴퓨터정보학과 학사

2009년~현재 고려대학교 컴퓨터정보학과 석사과정

<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

이 상 우 (Sang-woo Lee) 준회원



2010년 고려대학교 컴퓨터정보학과 학사

2010년~현재 고려대학교 컴퓨터정보학과 석사과정

<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

김 명 섭 (Myung-sup Kim) 정회원

1998년 포항공과대학교 전자계산학과 학사

1998년~2000년 포항공과대학교 컴퓨터공학과 석사

2000년~2004년 포항공과대학교 컴퓨터공학과 박사

2004년~2006년 Post-Doc., Dept. of ECE, Univ. of Toronto, Canada

2006년~현재 고려대학교 컴퓨터정보학과 조교수

<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석, 멀티미디어 네트워크