동작형태 분석을 통한 Skype 응용 트래픽의 실시간 탐지 방법

준회원 이 상 우*, 이 현 신*, 정회원 최 미 정**, 종신회원 김 명 섭***

Real-time Identification of Skype Application Traffic using Behavior Analysis

Sang-woo Lee*, Hyun-shin Lee* *Associate Members*,
Mi-jung Choi** *Regular Member*, Myung-sup Kim*** *Lifelong Member*

요 약

최근 인터넷 사용자의 증가와 고속 네트워크 망을 통한 네트워크 트래픽의 급증으로 효율적인 네트워크 트래픽 관리의 필요성이 더욱 커졌다. 효율적인 트래픽 관리를 위해서는 응용 프로그램 별 트래픽 분류의 연구가 선행되어야 하며 이미 많은 기존 논문에서 응용레벨 트래픽 분류에 대한 다양한 알고리즘을 제시하고 있다. 하지만 P2P기반의 Skype응용에 대해서는 분석율이 떨어져 이에 대한 연구가 더 필요한 실정이다. 본 논문에서는 payload 시그니쳐 기반 분석, 기계학습 기반 분석 등 기존의 방법론에 의존하지 않고 Skype응용의 트래픽 특성을 분석해 사용자들의 {IP, port} 리스트를 추출하고, 이를 이용해 네트워크 내에 발생하는 Skype응용 프로그램의 트래픽을 정확하게 탐지하는 실시간 탐지 알고리즘을 제안한다. 제안된 방법론은 학내 네트워크에 적용하여 그 타당성을 검증 하였다.

Key Words : Skype Application, Real Time Traffic Classification, Traffic Pattern, P2P Application

ABSTRACT

As the number of Internet users and applications is increasing, the importance of application traffic classification is growing more and more for efficient network management. While a number of methods for traffic classification have been introduced, such as signature-based and machine learning-based methods, Skype application, which uses encrypted communication on its own P2P network, is known as one of the most difficult traffic to identify. In this paper we propose a novel method to identify Skype application traffic on the fly. The main idea is to setup a list of Skype host information {IP, port} by examining the packets generated in the Skype login process and utilizes the list to identify other Skype traffic. By implementing the identification system and deploying it on our campus network, we proved the performance and feasibility of the proposed method.

Ⅰ. 서 론

최근 인터넷 사용자의 증가와 고속 네트워크의 보급으로 네트워크 트래픽이 급증하였다. 이것은 단순히 WWW, FTP, SMTP, DNS와 같은 전통적인 인터넷 서비스뿐만 아니라 멀티미디어, P2P (peer-to-peer),

※ 본 연구는 2009년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(2009-0090455).
* 고려대학교 컴퓨터정보학과 석사과정 ({sangwoo_lee, hyunshin-lee}@korea.ac.kr)
** 강원대학교 컴퓨터과학과 조교수 (mjchoi@kangwon.ac.kr)
*** 고려대학교 컴퓨터정보학과 조교수 (tmskim@korea.ac.kr)
논문번호 : KICS2010-04-162, 접수일자 : 2009년 4월 8일, 최종논문접수일자 : 2011년 1월 13일

게임, 스마트 폰 응용 등의 다양한 서비스의 증가로 인한 부분도 큰 몫을 차지하기 때문이다. 이에 따른 트래픽이 급증함에 따라 효과적인 네트워크 관리를 위해 트래픽 모니터링 및 분석의 중요성이 커지고 있다.

효과적인 네트워크 관리를 위해 선행되어야 할 것은 해당 트래픽이 어떤 응용 또는 어떤 서비스에서 발생 되었는가를 판별 하는 것이다. 이미 많은 기존 논문에서 응용레벨 트래픽을 분류에 대한 다양한 알고리즘을 제시하였지만 P2P기반의 Skype 응용[1]에 대해서는 분석율이 떨어져 이에 대한 연구가 더 필요한 실정이다.

본 논문은 payload 시그니쳐 기반 분석, machine learning 기반 분석 등 기존의 방법론에 의존하지 않고 Skype응용의 트래픽 특성을 분석해 해당 응용 트래픽을 정확하게 탐지할 수 있는 알고리즘과 실시간 분석 시스템을 제안하는 것을 목적으로 한다. Skype 응용은 P2P 방식의 메신저로써 사용자간 채팅, 음성통화, 화상통화, 전화 교환망을 통한 일반 전화, 파일전송 등의 기능을 제공한다. 안정적인 통화품질 서비스 제공과 일반 전화망에 비교해 저렴한 가격은 오늘날 전 세계적으로 가장 많이 사용하는 메신저로 만들었다. 하지만 엔터프라이즈 네트워크 관리자 입장에서 보면 Skype의 트래픽들은 기본적으로 암호화[2]-[4]가 되어있고, 해당 응용 설치 시 동적 포트 번호가 할당되고, 일반적인 프로토콜을 사용하지 않아[5],[6] Skype의 트래픽을 정확하게 탐지하는 것은 일반적인 분석방법론으로는 불가능하다. 하지만 Skype 응용 설치 때마다 달라지는 클라이언트 포트를 알아 낼 수 있다면 대부분의 트래픽을 쉽게 분류 할 수 있으며 각 호스트의 응용 사용시간 측정 및 제어 등 다양한 방면에서 활용이 가능하다.

본 논문에서는 Skype 응용 설치과정에서 동적으로 할당되는 클라이언트의 Skype 포트번호를 알아내기 위해 Skype 응용의 로그인 과정을 분석하였다. 로그인 과정을 각 단계별로 구분하여 분석하고 각 단계에서 발생하는 패킷의 내용을 조사함으로써 엔터프라이즈 네트워크 내의 Skype 사용자(호스트) 각각의 동적 포트번호{IP, port}를 추출 할 수 있었다. 또한 추출된 사용자들의 리스트를 내부 Skype 사용자와 연관되어 트래픽을 발생시키는 외부 호스트의 {IP, port}를 추출 하였다. 이렇게 추출된 내부 및 외부 사용자의 {IP, port} 리스트를 기반으로 Skype 트래픽을 탐지하는 탐지 모듈을 개발하여 적용한 결과 신뢰성 있는 결과를 얻어 낼 수 있었다.

본 논문에서 정의하는 Skype 응용 트래픽은 해당 응용 (Skype.exe프로세스)에서 발생하는 모든 트래픽으로 정의한다. 여기에는 Skype 고유의 응용 프로토콜을 사용하는 트래픽뿐만 아니라 HTTP, HTTPS 트래픽도 포함된다. Skype 트래픽의 탐지는 대상 네트워크 링크로부터 수집되는 모든 트래픽을 입력으로 하여 Skype응용 트래픽만 탐지하여 분류해 내는 것을 목표로 한다. 본 논문에서는 학내 네트워크의 인터넷 링크를 트래픽 수집 지점으로 설정하였다.

본 논문은 다음과 같은 순서로 구성되어 있다. 2장에서는 관련연구를 기술하고, 3장 에서는 본 논문에서 사용한 트래픽 수집 환경 및 검증 방법론을 기술한다. 4장 에서는 Skype 의 로그인 단계들을 도식화하여 순차적으로 보여주며, 5장 에서는 자세한 탐지 알고리즘을 설명한다. 6장 에서는 실제 학내 망에 탐지 시스템을 적용 후 검증결과를 보여주며 마지막으로 7장 에서는 결론 및 향후 연구과제를 기술한다.

Ⅱ. 관련연구

이미 기존의 많은 연구들에서 응용 트래픽 분류를 위한 DPI 기반, 기계학습 기반, 패킷 사이즈와 포트 기반 방식들이 제안되고 있지만 Skype응용에 대해서는 분류의 정확성이 높지 않거나 향후 연구로 남겨두고 있다[3,4,7,8].

[3],[4]에서는 트래픽 분석을 위해 payload 기반 분류 방법론을 제안했으나 Skype응용에 대해서는 그 데이터가 암호화 되어있어 분류가 어렵다고 기술하고 있다. 물론 payload기반 분류 방법은 데이터 부분이 암호화 되어있으면 분류가 불가능하겠지만, Skype 응용의 경우는 UDP패킷이나 Skype네트워크 상에 로그인 과정에서 발생하는 패킷에는 시그니쳐가 존재하며 이것은 분명 트래픽 분류 시 중요한 단서가 될 수 있다.

[7],[8]에서는 기계학습 기법을 이용해 P2P응용 프로그램을 분류하는 방법을 제안하고 있다. 하지만 제안하는 방법론은 별도의 학습을 위한 데이터 셋이 필요하며 Skype응용에 한해서는 분류 정확성이 높지 않은 결과를 보여주고 있다. 본 논문에서 제안하는 방법론은 별도의 학습을 위한 데이터 셋이 필요하지 않으며 동적 포트번호를 추출과 동시에 정확하게 분류 할 수 있는 장점을 지니고 있다.

[9]에서는 패킷 사이즈 분포를 이용한 분류 방법론을 제안하고 있다. 하지만 Skype 구 버전인 3.0에 대해 적용하였고, 동적인 Skype 호스트 각각의 포트를 추출하기 어렵다는 단점을 가지고 있다. 본 논문에서도 일부 패킷 사이즈를 이용한 분류 방법을 제안하고 있지만 Skype 최신 버전인 4.1 에 대해 적용이 가능하며 Skype 호스트 각각의 포트를 추출함에 따라 엔터프라이즈 네트워크 내에 발생하는 Skype응용의 트래픽에 대해 다양한 측정이 가능하다.

본 논문에서는 실시간 Skype응용의 트래픽 분류 시스템을 위해 Skype사용자 각각의 {IP, port}를 추출함과 동시에 분류 작업을 수행하며 분류의 정확성과 속도를 목표로 하였다. TCP flow의 경우 상위 5개 패킷까지, UDP flow의 경우 상위 2개의 패킷까지 조사하며 조사하는 패킷의 개수가 더 이상 증가하면 해당 flow는 Skype응용의 flow가 아님으로 판단하여 시스템 부하를 줄이도록 하였다. 또한 암호화된 패킷에 대해 복호화 작업을 거치지 않아 개인 프라이버시 문제에 대처 할 수 있다.

Ⅲ. 트래픽 수집 및 검증 시스템

본 장에서는 Skype응용 트래픽을 탐지하기 위해 구성한 트래픽 수집 시스템 및 검증 시스템을 기술한다. 트래픽 수집은 본 연구실에서 개발한 실시간 트래픽 모니터링 시스템인 KU-MON[10]을 외부 인터넷 망과 연결되는 라우터에 설치하여 실시간으로 패킷을 수집하였다. 그리고 본 논문에서 제안하는 방법론의 검증을 위한 정답지(Ground Truth Data) 생성 방법론으로 종단 호스트에 TMA[11](Traffic Measurement Agent)를 설치하는 agent 활용 기법을 사용하였다.

3.1 트래픽 수집 시스템

트래픽 수집은 KU-MON 을 외부 인터넷 망과 연결되어있는 라우터에 설치하여 수집하였다. 수집된 패킷은 flow단위로 그룹화 되는데, 본 논문에서는 flow를 패킷 헤더의 5-tuple (Source IP, Source port, Destination IP, Destination port, Protocol)정보를 공유하는 양방향(Twoway flow)패킷들의 집합으로 정의하였다. 그림 1은 본 논문에서 사용한 KU-MON 기반 트래픽 수집 및 검증 시스템을 나타낸 것이다. 그림1에서와 같이 라우터에 연결된 TCS는 미러링을 통해 패킷들을 수집하여 실시간으로 flow를 생성한다. 생성된 flow는 TAS와 TMS로 각각 전송된다. 본 논문에서 제시하는 동작형태 분석기반 Skype트래픽 탐지 시스템은 TCS와 TAS상에서 동작하며 Real-time SIS라 정의한다. Flow의 생성은 실시간으로 이루어지며, 생성과정에서 flow 단위로 Skype응용인지 아닌지를 탐지한다.

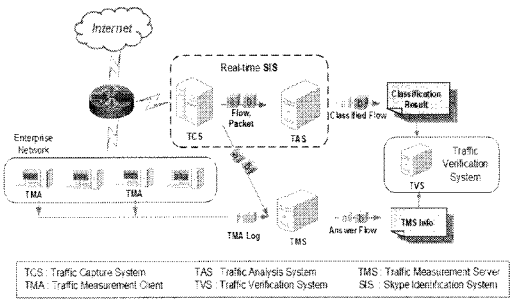


그림 1. 트래픽 수집 및 검증 시스템
Fig. 1. Traffic Collection, Verification System

3.2 TMA 검증 시스템

본 논문에서 사용된 검증 시스템은 네트워크 내의 종단 호스트에 설치된 TMA를 이용하여 Skype 트래픽 정답지를 생성하고 이를 바탕으로 분류결과를 검증하는 방법을 사용하였다. 그림 1에서 종단 호스트에 설치된 TMA는 해당 호스트의 현재 활성화된 소켓 정보를 토대로 TMA정보(Process Name, Source IP, Source port, Destination IP, Destination port, Protocol)를 추출하여 TMS로 전송한다. TMS는 이를 이용하여 TCS로부터 전달받은 flow 데이터와 비교하여 Skype 응용에 대한 정답지 flow를 생성한다. TVS(Traffic Verification System)에서는 TAS에서의 탐지 결과와 정답지 flow의 비교를 통해 탐지결과의 정확도를 측정한다.

Ⅳ. Skype 로그인 단계

본 논문에서는 Skype 응용을 사용하는 호스트 리스트를 구축하고 이를 기반으로 Skype 트래픽을 탐지하는 방법을 제안한다. Skype응용의 로그인 단계에 나타나는 동작특성 분석을 통하여 Skype 호스트 리스트의 구축하였다. 즉, 로그인 단계 트래픽 분석을 통하여 Skype 응용을 사용하는 호스트를 탐지하고, 그 호스트와 통신하는 상대 호스트를 탐지하는 과정으로 Skype 호스트 리스트를 구축한다.

그림 2는 Skype응용의 로그인 단계를 나타낸 그림이다. 네트워크상에 그림 2와 같은 패턴의 트래픽

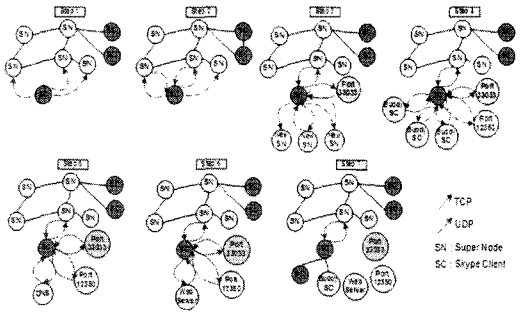


그림 2. Skype 응용의 Login 단계
Fig. 2. Skype Application Login Step

이 나타났을 때 해당 호스트가 Skype 네트워크상에 로그인 하였다는 것을 추측 할 수 있다. 다음은 그림 2에서와 같이 Skype로그인 단계를 7단계로 나누어 기술한다.

1) Step_1-SC(Skype Client)에 저장되어 있던 SN(Super Node)[12]리스트를 기반으로 SC가 SN들에게 UDP패킷을 전송한다.
2) Step_2-SC에서 UDP패킷을 보낸 SN으로부터 UDP 응답 패킷을 받는다.
3) Step_3-응답 패킷이 온 SN에게 TCP 연결을 맺는다. 여러 SN들로부터 응답이 온 경우 응답 패킷을 보낸 첫 번째 또는 두 번째 SN을 선택하여 TCP 연결을 맺는다. 이 TCP 연결을 통하여 SC는 새로운 SN 리스트를 전송받는다고 추정된다. 리스트를 받음과 동시에 UDP 패킷을 새로운SN에게 전송하고 응답 패킷을 기다린다. SC는 UDP응답 패킷이 도착한 SN들에 대하여 SN 리스트를 갱신한다. 이 때 목적지 포트번호가 33033인 새로운 노드로 TCP 연결도 동시에 이루어진다.
4) Step_4-SC가 목적지 포트번호가 12350인 노드로 UDP패킷을 전송한다. 그리고 SC와 buddy (Skype친구등록)관계인 SC들에게 UDP패킷을 전송하며 만약 프로필 변경, 사진 변경 등 갱신 정보가 있을 때 buddy 관계의 SC와 TCP 연결을 맺어 갱신정보를 전송 받는다.
5) Step_5-DNS 서버로 DNS 쿼리를 전송하여 Skype 웹 서버의 IP를 받아온다. 이 단계에서는 목적지 포트번호가 12350 인 노드로TCP 연결을 맺는 경우와 맺지 않는 경우의 두 가지 경우가 생긴다. SC가 이전에 로그인하였던 호스트에서 다시 로그인 하였을 경우는 목적지 포트번호가 12350인 노드와 TCP 연결을 맺지 않고, SC 가 이전과 다른 새로운 호스트에서 로그인할 경우목적지 포트번호가 12350인 노드와 TCP 연결을 맺게 된다.
6) Step_6-DNS서버로부터 받은 IP인 웹 서버와 TCP 연결을 맺고 버전을 체크한다.
7) Step_7-목적지 포트번호가 33033인 노드와 접속을 해제하며, step_5에서 목적지 포트번호가 12350인 노드와 TCP 연결을 맺었을 경우 이 단계에서 접속을 해제한다.
8) Step_7 이후-buddy SC와 통신, 그 외 SC와 통신, 로그아웃 할 경우 step_3단계에서SN과 맺었던 TCP 연결을 끊게 된다.

본 논문에서는 SC 및 SN의 Skype 포트번호 추출을 위해 Step_4의 목적지 포트번호 12350인 노드로 트래픽을 전송할 때 발생하는 UDP 패킷의 크기와 페이로드 내용을 분석한다. 또한 로그인 단계가 끝난Step_7이후에서도 구축된 Skype 호스트 리스트와의 상관관계를 통하여 다른 호스트의 {IP, port} 또한 추출이 가능하다.

V. 분류 알고리즘

본 논문에서 제시하는 SIS시스템은 네트워크 링크로부터 패킷을 받아 실시간으로 flow를 생성 및 갱신하면서 해당 flow가 Skype flow 인지 판단한다. TCP flow의 경우 최대 5개 패킷까지, UDP flow의 경우 최대 2개 패킷까지 조사하기 때문에 트래픽의 발생초기에 빠른 속도로 탐지 할 수 있다. 또한 추출된 Skype응용 호스트의 {IP, port} 리스트를 구성함으로써 패킷의 내용을 살펴보지 않고 flow의 헤더 정보만으로 탐지가 가능하다.

5.1 Skype응용 트래픽 탐지 알고리즘

본 논문에서 제안하는 Skype 응용 트래픽 탐지 알고리즘의 핵심은 트래픽 flow의 초기 몇 개 패킷의 DPI 분석을 통하여 탐지하고 이를 바탕으로 SC가 설치된 호스트의 {IP, port} 리스트를 구축함으로써 다른 SC로부터 발생하는 Skype flow를 쉽게 탐지하는 것이다. {IP, port} 리스트는 항상 최신의 정보가 유지되도록 다양한 aging 기법을 이용하여 관리된다. 그림 3은 본 논문에서 제시하는 Skype 트래픽 탐지 알고리즘의 순서도이다. 먼저 패킷이 캡쳐되면 해당 패킷을 기반으로 flow 정보를 생성 또는 이미 생성된 flow의 경우 flow정보를 갱신한다. 해당 flow가 이미 Skype flow인지 아닌지 이미

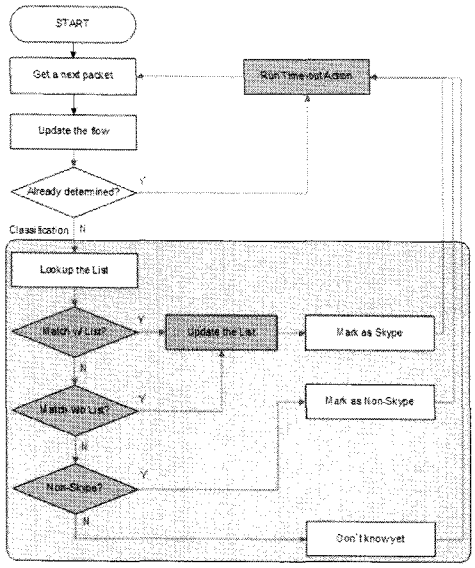


그림 3. 탐지 알고리즘 순서도
Fig. 3. Identification Algorithm Flow Chart

결정 된 상태라면 더 이상 검사를 수행하지 않고 다음 패킷을 기다린다. 하지만 아직 결정되지 않은 flow라면 Classification 단계에 들어서게 된다. Classification 단계에서는 해당 flow가 Skype응용인지(Mark as Skype) 아닌지 (Mark as Non-skype) 결정을 내리거나, 더 많은 패킷들을 살펴봐야 한다고(Don't know yet)판단을 내리게 된다. 또한 해당 flow가 Skype응용의 flow로 판단 시 새롭게 생성된 호스트의 {IP, port}정보를 리스트에 추가하고, 해당 트래픽이 나타난 시점(Last Usage)을 기록한다. 본 탐지 알고리즘에서는 TCP flow경우 5개의 패킷까지 조사하며 UDP flow경우 2개의 패킷까지 조사한다. 패킷의 최소 필요 개수를 만족 시키지 못 할 경우 다음 패킷까지 조사를 수행해야 함으로 판단을 보류(Don't know yet) 하고 시스템은 다음 패킷을 기다린다.

5.2 탐지 알고리즘에서 구축하는 {IP, Port}리스트

제안된 알고리즘은 Skype트래픽의 탐지과정에서 Skype 응용 호스트의 {IP, port}를 추출하며 리스트를 구성한다. 리스트에는 현재 Skype를 사용하는 {IP, port} 정보가 저장되며 호스트당 1개의 {IP, port} 정보가 저장된다. 리스트의 각 레코드는 {IP, port} 정보 외에 그 정보의 Skype 트래픽이 발생한 마지막 시간을 저장하는 last_usage 값을 가진다. 리스트에서 해당 {IP, port} 정보의 트래픽이 2일동안 나타나지 않는 경우 해당 호스트에서는 더 이상 Skype 응용이 실행되고 있지 않다고 판단하고 해당 정보를 리스트에서 삭제한다. 이 작업은 그림 3에서 [Run Time-out Action] 모듈에서 수행한다. 이는 탐지 시스템의 부하를 줄이고 항상 최신의 {IP, port} 정보를 유지할 수 있게 한다. 리스트는 탐지 시스템이 실행된 시점부터 주기적으로 백업 되며 탐지 시스템의 작동이 중단되더라도 백업된 리스트를 불러들임에 따라 해당 정보를 계속적으로 유지할 수 있다.

5.3 탐지 모듈

그림 3의 각 모듈의 탐지 조건들을 기술하기 위하여 다음 표 1와 같은 표기 방식을 사용한다.

다음 표 2는 각 탐지 모듈에서의 살펴보는 패킷의 개수와 프로토콜, 살펴볼 조건을 나타낸다. ①번 조건은 TCP/UDP 모든 flow에 대하여 payload의 조사 없이 {IP, port} 만으로 수행되고, ⑤번의 경우는 UDP flow의 포트번호와 2번째 패킷의 payload를 조사한다는 것이다. 표 2에서도 알 수 있듯이 제안된 탐지 알고리즘은 TCP의 경우 5개의 payload 패킷을 참조, UDP의 경우 최대 2개의 패킷까지만 참조한다. TCP의 경우 3번째, 4번째, 5번째 패킷에서, UDP의 경우 1번째, 2번째에서 고유의 패킷 사이즈를 가지는 패턴을 보이기 때문이다.

그림 3에서 보는 바와 같이 Skype응용 트래픽 탐지 시스템은 [Match w/ List], [Match w/o List]

표 1. 모듈의 조건 표기 방식
Table. 1. Module Representation

Symbol	Description
src	flow 정보에서 source {IP, port} 정보
dst	flow 정보에서 destination {IP, port} 정보
L	Skype호스트의 {IP, port} 리스트
P[ ]	패킷 payload의 octet 배열 (16진수)
|Pn|	Flow 내의 payload가 있는 n번째 패킷의 크기
a → L	A를 L에 삽입함

표 2. 모듈의 조건 별 필요 요소
Table 2. Module Condition needs

Payload Packet Nubmer		-	1	2	3	4	5
Protocol	Condition						
TCP/UDP	{IP, port}	①					
TCP	{IP}				②		
	{port}		③				
UDP	{port}		④	⑤			
	-		⑥				

두 가지 모듈로 구성되어있다. 전자는 Skype flow 탐지에 리스트의 정보를 참조하는 것이고 후자는 참조하지 않는 방법이다. 다음 그림 4와 5는 각 탐지 모듈의 탐지 조건 및 Skype 호스트 리스트 삽입 내용을 나타낸다.

패킷의 개수와 프로토콜을 살펴보지 않는 ①번 조건은 그림4의 [Match w/ List] 모듈에 속해 있으며 flow의 IP, port 정보와 추출된 리스트만으로 판별하며 src(IP port)가 추출된 리스트에 있을 시 dst(IP port)를 리스트에 삽입함을 의미한다. ①번과 같은 조건은 해당포트가 Skype응용 포트라면 그와 연관되어 트래픽을 발생하는 다른 호스트의 포트번호도 Skype 응용의 포트번호이라는 가정하에 고안 되었다. ②번 조건의 경우도 [Match w/ List] 모듈에 속해 있으며 flow의 프로토콜이 TCP이며 3, 4, 5 번째 패킷의 크기를 통해 해당 flow가 Skype응용의 트래픽인지 판단한다.

③, ④, ⑤, ⑥번 조건은 [Match w/o List] 모듈에 속해 있으며 리스트 참조 없이 flow의 패킷의 크기와 포트번호, payload 시그니쳐 등을 통해 판단하며 최대 2개의 패킷까지만 조사한다. ③번 조건의 경우 목적지 포트번호가 33033이고 payload 의 크기가 5 bytes, payload 시그니쳐가 16 03 01 00 00 일 경우 이는

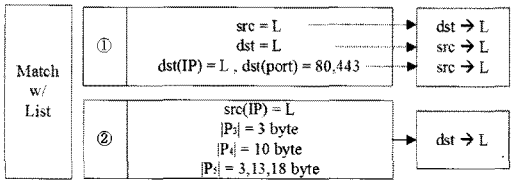


그림 4. 리스트를 사용하지 않는 매칭
Fig. 4. Match w/ List

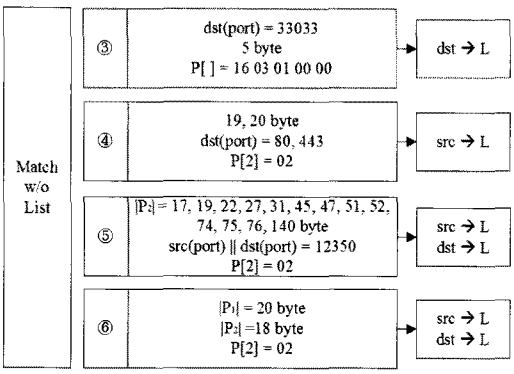


그림 5. 리스트를 사용하여 매칭
Fig. 5. Match w/o List

Skype로그인 과정에서 나타나는 패턴으로 판단 가능하다. 목적지 호스트가 SN이므로 이는 Skype네트워크의 또 다른 사용자이기 때문에 리스트에 삽입한다. ⑤번의 조건의 경우 또한 로그인 과정에서 나타나는 패턴이며 이 또한 다른 사용자 이기 때문에 리스트에 삽입한다. ④번과 ⑥번의 조건은 해당응용을 사용시 주기적으로 다른 새로운 노드와 계속적으로 UDP패킷을 주고 받는데 패킷의 사이즈와 payload 시그니쳐가 존재하며 이는 해당응용이 주기적으로 새로운 SN 노드들을 찾는 패킷으로 추정된다.

[Non-Skype] 모듈에서는 Skype flow가 아닌 flow를 최종 판단한다. TCP의 경우 6개 이상, UDP의 경우 3개 이상의 payload 패킷이 발생했을 경우 Skype가 아니라고 판단한다. 또한 패킷 개수의 한계를 넘지 않더라도 flow가 끝나면 Skype가 아니다.

VI. 적용 및 검증 결과

표 3은 2010년 2월26일 오전 00시 00분부터 3월 1일 오후 24시까지 4일 동안 고려대학교 캠퍼스 네트워크에서 수집된 트래픽 트레이스이다. 해당 트레이스가 수집되는 동안 7,583개의 호스트가 네트워크를 사용하였으며 약 2.5 TB의 트래픽을 발생시켰다.

다음 표 4는 TMA를 통해 생성된 정답지 데이터의 트레이스 이다.

본 논문에서의 실험은 추출된 리스트가 전혀 없는 상태에서 측정 되었으며 해당 트레이스 전체에서 Skype응용의 트래픽 만을 분류를 하는 것을 목적으로 하였다. 검증은 리스트가 어느 정도 모였다고 판단되는 2월26일 01시 30분부터 하였다. 이유는 해당 분류기는 이전의 리스트를 기반으로 동적인 리스트를 구성하는데 리스트를 초기화 시킨다면 트레이스 이전에 로그인한 호스트에 대해서는 분석이 불가능하기 때문이다. 표 5는 해당 알고리즘에 의해 분류된 Skype응용의 Flow, Packet, Byte 및 분류된 트래픽에서의 비율을 나타낸다.

표 3. 수집된 트래픽 트레이스
Table 3. Traffic Trace

10/02/26(00:00)~ 10/03/01(24:00)	TCP	UDP	Total
# of flow	56 x 10^6	31 x 10^6	87 x 10^6
# of packets	1,622 x 10^6	540 x 10^6	4,126 x 10^6
Bytes	1083 GB	239 GB	2,625 GB
Hosts	7,583		

표 4. 정답지 데이터
Table 4. Ground truth data

10/02/26(00:00)~ 10/03/01(24:00)	TCP	UDP	Total
# of flow	6 x 10^6	7 x 10^6	14 x 10^6
# of packets	302 x 10^6	263 x 10^6	565 x 10^6
Bytes	104 GB	119 GB	224 GB
Hosts	1472		

표 5. 분석된 Skype 응용의 flow, packet, byte
Table 5. Identified Skype Application flow, packet, byte

	TCP	UDP	Total
# of flow (%)	46,535 (46)	54,543 (54)	101,078 (100)
# of packets (%)	414,821 (40)	599,336 (60)	1,014,157 (100)
Bytes (%)	34 MB (17)	164MB (83)	198MB (100)
Byte Completeness	0.003 %	0.068 %	0.007 %

분류된 결과를 살펴보면 Skype응용이 UDP 프로토콜을 통해 대부분의 데이터를 전송한다는 것을 알 수 있으며 byte, packet, flow 순으로 많다. 이는 Skype응용이 UDP 프로토콜을 기반으로 동작 한다는 것을 보여준다. 결과의 정확도는 표 6에서 보는 바와 같이 Precision, Recall의 값이 모두 100%로 오탐없이 정확하게 Skype 트래픽을 탐지 할 수 있음을 알 수 있다.

TP(True Positive)는 해당응용의 트래픽을 분류알고리즘에서 정확하게 탐지한 개수를 나타내고, FP(False Positive)는 해당응용의 트래픽을 다른 응용의 트래픽으로 잘못 탐지한 개수를 나타낸다. FN(False Negative)는 해당응용의 트래픽이나 이를 탐지 못한 개수를 나타낸다.

그림 6은 분류 단계에서 각 그림 4와 그림 5 모

표 6. 탐지결과의 정확도
Table 6. Identified Result Accuracy

	Value
TP (# of flow)	11828
FP (# of flow)	0
FN (# of flow)	0
Precision (%)	100.0
Recall (%)	100.0

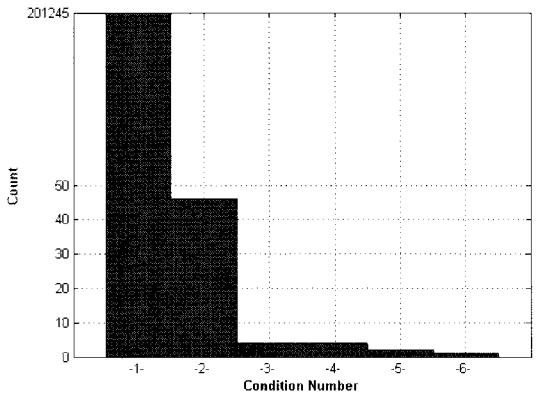


그림 6. 분류에 사용된 조건 별 사용빈도
Fig. 6. Usage of Module's Condition

듈의 조건 별 사용빈도수를 나타내는 그래프 이다.

대부분의 트래픽들이 [Match w/ List] 모듈의 ① 번 조건에 의해 분석됨을 알 수 있다. 이것은 해당 응용이 호스트에 설치될 시 랜덤 포트를 할당 받으나, 설치된 후에는 할당 받은 포트를 계속적으로 사용한다는 것을 의미한다. 또한 추출된 호스트들의 리스트들이 신뢰성 있다는 것을 의미하며, 리스트가 구성된 후에는 해당 flow의 패킷 정보를 보지 않고서도 헤더정보만으로 분류 할 수 있다는 것을 의미한다. 다음 표 7은 트레이스에서 분류된 학내 망 내부 호스트(Inside)의 수와 그와 연관된 망 외부 호스트(Outside)의 수를 보여준다.

이는 엔터프라이즈 네트워크 내부의 호스트 하나가 평균 1100개의 외부 호스트와 트래픽을 발생 시켰다는 것을 의미하며 다른 P2P응용에 비하여 매우 많은 Connection을 발생 시킨다는 것을 알 수 있다.

표 8은 26일 평균 사용시간, flow, packet, byte

표 7. 수집된 네트워크 내부, 외부 사용자 수
Table 7. User # of Inside, Outside

	Inside	Outside	Total
Count	40	69282	69322

표 8. Skype 사용시간 및 flow, packet, byte 발생량
Table 8. Skype usage time, host's flow, packet, byte

	Value
평균 사용자수	24
총 사용시간 (avg)	6627 (276.1) min
flow (avg)	24,739 (1030)
packet (avg)	331,519 (13,813)
byte (avg)	134.1 (5.5) MB

를 나타낸다. 각 Skype응용 호스트당 평균 총 4.6 시간을 사용하였고, 대부분의 트래픽이 해당 날에 의해 발생되었다는 것을 의미한다.

그림 7은 각 호스트당 접속 유지 시간 별 횟수 그래프이다. 위 그래프를 통해 대부분의 호스트는 2 시간 미만의 접속 시간을 유지하며, 표8 의 평균 사용시간과 비교해 보았을 때 호스트들은 Skype네트워크 상에 자주 로그인 한다는 것을 알 수 있다.

다음 그림 8은 리스트초기화 후 3시간30분 동안의 FN 변화 그래프이다.

본 논문에서의 실험은 리스트가 충분히 모인 1시간 30분 이후 검증이 이루어졌다. 그림8의 FN변화 그래프는 리스트초기화 이후 분류기 에서 사용자 파악에 얼마나 시간이 걸리는 지를 나타낸다. 초기 추출된 사용자가 없더라도 본 분류기는 1시간 30분 만에 사용자 추출에 성공하고 그 이후에는 FN이 나타나지 않는 것을 확인 할 수 있다.

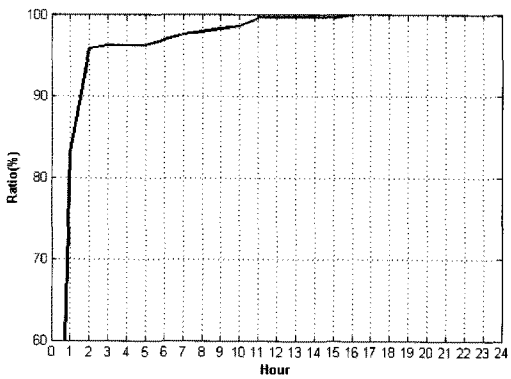


그림 7. Skype 호스트당 접속 유지시간 분포
Fig. 7. Distribution of Skype host connect time

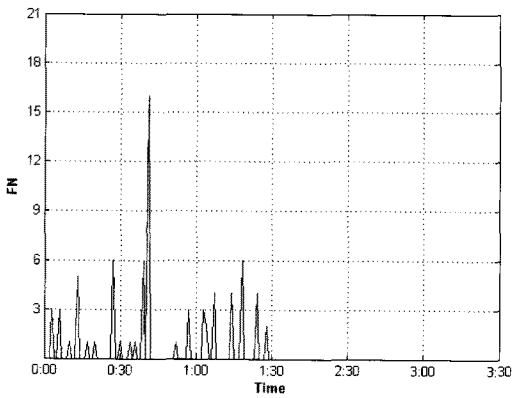


그림 8. 시간에 따른 FN의 변화
Fig. 8. FN changes

다음 그림 9는 하루 24시간(1440분) 동안 분류기의 Processing Time(Second), 학내 네트워크 bps(Bit Per Second)을 나타낸 그래프이다. 분류기의 Processing Time은 추출된 사용자의 리스트가 없을 때부터 측정되었다. 대체적으로 1초 이내로 측정되며 네트워크 사용량이 증가함에 따라 Processing Time이 증가하지만 급작스러운 트래픽 증가에 robust한 것을 보여준다.

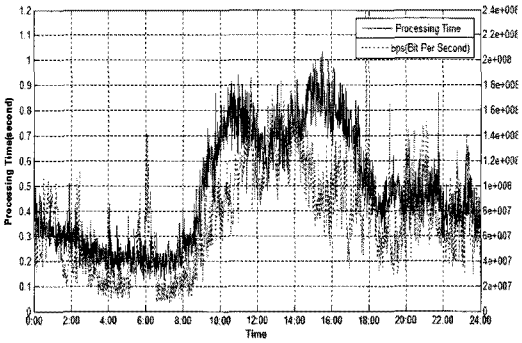


그림 9. 분류기 Processing Time 및 트레이스 bps
Fig. 9. Identification system Processing Time, Trace bps

Ⅶ. 결론 및 향후 연구

본 논문에서는 기존의 연구들에서 응용 트래픽 분류를 위한 DPI 기반, 기계학습 기반, 패킷 사이즈와 포트 기반 방법론들에서 탐지하지 못하는 Skype 응용의 트래픽을 탐지하는 방법론을 제시 하였다. 제안된 Skype응용 트래픽 탐지 알고리즘은 해당 응용의 로그인 과정을 분석해 응용의 고유 패턴을 이용해 {IP, port}리스트를 추출하고 이를 바탕으로 다른 Skype응용 트래픽을 탐지하였다. 향후 연구로 리스트를 보유하고 있지 않은 상태에서 사용자 파악에 드는 시간을 줄이는 탐지알고리즘의 성능향상과 다른 응용프로그램들의 패턴파악에 따른 트래픽의 기능별 탐지에 초점을 두고 있다.

참 고 문 헌

[1] Skype, Web Site, http://www.skype.com.

[2] S. A. Baset, H. Schulzrinne, "An analysis of the skype peer-to-peer internet telephony protocol", *In Proceedings of IEEE INFOCOM'06*, pp.1-11.Barcelona, Spain, Apr. 2006,

[3] Risso, F., Baldi, M., Morandi, O., Baldini, A.,

Monclus, P., "Lightweight, payload-based traffic classification: An experimental evaluation", *In Proceedings of IEEE International Conference on Communications* ICC pp. 5869-5875, 2008.

[4] N. Cascarano, L. Ciminiera, F. Risso, "Improving Cost and Accuracy of DPI Traffic Classifiers", *25th ACM Symposium On Applied Computing(SAC 2010)*, pp.643-648, Sierre, Switzerland, March. 2010.

[5] Dario Rossi, Marco Mellia, Michela Meo, "Understanding Skype signaling", *Computer Networks 2009*, pp.130-140, Vol.53 No.2, 2009.

[6] Raffaele Bolla, Riccardo Rapuzzi, and Michele Sciuto, "Monitoring and Classification of Teletraffic in P2P Environment" *Proc. of the 2006 Australian Telecommunication Networks and application Conference*, Melbourne, Australia, Dec. 4-6, 2006.

[7] H.Liu, W.Feng, Y.Huang, X.Li, "A peer-to-peer traffic identification method using machine learning", *in International Conference on Networking, Architecture, and Storage, NAS*, pp.155-160, July. 29-31. 2007.

[8] Zhu Li, Ruixi Yuan, and Xiaohong Guan, "Accurate Classification of the Internet Traffic Based on the SVM Method" *Proc. of the IEEE International Conference on Communications*, pp. 1373-1378, Glasgow, Scotland, Jun. 24-28, 2006.

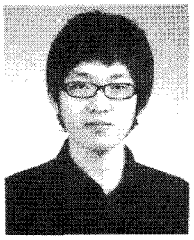
[9] Ying-Dar Lin, Chun-Nan Lu, Yuan-Cheng Lai, Wei-Hao Peng, Po-Ching Lin, "Application classification using packet size distribution and port association", *Journal of Network and Computer Applications*, Vol.32, pp. 1023-103, Issue 5, Sep. 2009.

[10] KUMON, Web Site, http://kumon.korea.ac.kr

[11] 윤성호, 노현구, 김명섭, "TMA(Traffic Measurement Agent)를 이용한 인터넷 응용 트래픽 분류", *통신학회 하계종합학술발표회*, pp.618, 라마다플라자호텔, Jul. 2-4, 2008.

[12] Yanfeng Yu, Dadi Liu, Jian Li, Changxiang Shen, "Traffic Identification and Overlay Measurement of Skype", *Computational Intelligence and Security*, 2006 International Conference, pp.1043-1048, Vol.2, Nov. 3-6, 2006.

이 상 우 (Sang-woo Lee) 준회원

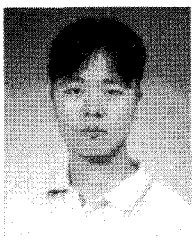


2010년 고려대학교 컴퓨터정보학과 학사 졸업

2010년~현재 고려대학교 컴퓨터 정보학과 석사과정

<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

이 현 신 (Hyun-shin Lee) 준회원



2009년 고려대학교 수학과 학사

2009년~현재 고려대학교 컴퓨터정보학과 석사과정

<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

최 미 정 (Mi-jung Choi) 정회원

1998년 이화여자대학교 컴퓨터공학과 학사

2000년 포항공과대학교 컴퓨터공학과 석사

2004년 포항공과대학교 컴퓨터공학 박사

2004년 3월~2004년 9월 포항공대 컴퓨터공학과 박사후 연구원

2004년 10월~2005년 9월 프랑스 INRIA 연구소 박사후 연구원

2005년 11월~2006년 10월 캐나다 워터루대학 컴퓨터과학부 박사후 연구원

2006년 11월~2008년 8월 포항공대 컴퓨터공학과 연구조교수

2008년 8월~현재 강원대학교 컴퓨터과학과 조교수

<관심분야> 네트워크 및 서비스 관리, 트래픽 측정 및 분석, 미래 인터넷, M2M 네트워크 및 서비스 관리

김 명 섭 (Myung-sup Kim) 종신회원

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