플로우 상관관계를 통한 인터넷 응용 트래픽 분석의 성능 향상에 관한 연구

준회원 윤 성 호*, 종신회원 김 명 섭*°

A Study of Performance Improvement of Internet Application Traffic Identification using Flow Correlation

Sung-ho Yoon* *Associate Member*, Myung-sup Kim*° *Lifelong Member*

요 약

인터넷의 대중화로 네트워크 트래픽은 지속적으로 증가하고 복잡해지고 있다. 따라서 네트워크 자원의 효율적인 사용을 위한 응용 트래픽 분석의 중요성은 더욱 강조되고 있다. 본 논문에서는 기존 시그니쳐 기반의 인터넷 트래픽 분석 방법의 한계점을 극복하고 분석 결과의 성능 향상(분석률)을 위해 플로우 상관관계를 이용한 트래픽 분석 방법을 제안한다. 본 논문에서 제안하는 방법은 시그니쳐 기반 분석기의 결과를 입력 받아 분석된 플로우와 그렇지 않은 플로우들 간의 상관관계를 파악하고 이를 통해 분석되지 않은 플로우를 분석한다. 총 4가지(서버-클라이언트, 발생 시간, 호스트-호스트, 통계) 세부 분석 방법과 이를 통합한 플로우 상관관계 기반 분석기를 제안한다. 또한 실험과 검증을 통하여 플로우 상관관계 기반 응용 분석 방법의 타당성을 증명한다.

Key Words : Flow Correlation, Traffic Monitoring and Analysis, Traffic identification

ABSTRACT

As network traffic is dramatically increasing due to the popularization of Internet, the need for application traffic identification becomes important for the effective use of network resources. In this paper, we present an Internet application traffic identification method based on flow correlation to overcome limitation of signature-based identification methods and to improve performance (completeness) of it. The proposed method can identify unidentified flows from signature-based method using flow correlation between identified and unidentified flows. We propose four separate correlation methods such as Server-Client, Time, Host-Host, and Statistic correlation and describe a flow correlation-based identification system architecture which incorporates the four separate methods. Also we prove the feasibility and applicability of our proposed method by an acceptable experimental result.

Ⅰ. 서 론

최근 인터넷 사용자의 증가와 고속 네트워크의 보급으로 네트워크 트래픽이 급증하였다. 이는 단순한 WWW, FTP, E-mail 과 같은 전통적인 인터넷 서비스뿐만 아니라 멀티미디어 스트리밍, P2P(peer-to-peer) 파일 공유, 게임과 같은 다양한 서비스의 등장 때문이다. 인터넷 트래픽이 급증함에 따라 효율적인 네트워크 관리를 위한 트래픽 모니터링 및 분석의 중요성이 커지고 있다[1,2].

※ 본 연구는 2009년 정부(교육과학기술부)의 재원으로 한국연구재단(2009-0090455)의 지원을 받아 수행되었습니다.
* 고려대학교 컴퓨터정보학과({sungho_yoon, tmskim}@korea.ac.kr), (°: 교신저자)
논문번호 : KICS2011-01-014, 접수일자 : 2011년 1월 6일, 최종논문접수일자 : 2011년 5월 17일

인터넷 트래픽 분석은 다양한 목적으로 이루어질 수 있지만 본 논문에서는 분석 대상 네트워크의 트래픽을 수집하여 분류 기준(프로토콜, 응용, 타입 등)에 맞게 트래픽을 분류하고, 수량적으로 측정을 하는 것을 목적으로 한다. 본 논문에서는 트래픽을 발생시킨 응용을 분류 기준으로 사용한다. 이렇게 분석된 정보는 네트워크 관리 및 보안에 중요한 자료로 쓰인다. 특히, 효과적인 네트워크 자원 관리를 위해 특정 응용에서 발생하는 트래픽의 대역폭을 조절하거나 차단하기 위해서는 실시간 트래픽 분석이 반드시 선행되어야 한다.

최근 들어 정확한 트래픽 분석을 위한 많은 방법론이 제시되었다. 그 중 상업적으로 가장 많이 사용하는 방법은 시그니쳐 기반 분석 방법이다. 시그니쳐 기반 트래픽 분석 방법이란, 특정 응용이 발생하는 트래픽을 수집하여 해당 응용 트래픽에서만 나타나는 고유한 특징을 시그니쳐로 추출하고 분석 대상 트래픽과 시그니쳐의 비교를 통해 응용을 판단하는 방법이다. 대표적인 시그니쳐는 페이로드에 존재하는 고유한 문자열을 사용하는 페이로드 시그니쳐이다[3]. 최근에는 헤더 정보나 통계적 정보에서 응용 시그니쳐를 찾는 방법이 제안되고 있다[4,5].

시그니쳐 기반 분석 방법은 해당 응용의 시그니쳐를 정확히 추출하였을 경우, 매우 높은 성능(분석률과 정확도)을 보장 하지만 몇 가지 한계점[6,7]을 가진다. 첫째, 시그니쳐를 추출하지 못하는 응용은 분석할 수 없다. 시그니쳐를 추출하는 방법이 계속 발전되고 있지만, 원천적으로 시그니쳐 추출이 불가능한 경우가 존재한다. 예를 들어 페이로드가 없거나 암호화 되어 있는 경우에는 페이로드 시그니쳐를 생성할 수 없고, 데이터 서버를 공유하거나 암호화된 주소를 사용하는 응용의 경우에는 헤더 시그니쳐를 생성하지 못한다. 또한, 같은 엔진 기반의 응용이거나 같은 응용 레벨 프로토콜을 사용하는 경우 통계 시그니쳐를 추출할 수 없다.

둘째, 최신의 시그니쳐를 유지 및 관리하는 것은 어렵다. 응용은 사용자의 요구와 수요에 따라 자주 수정되고 변화된다. 따라서 최신 시그니쳐를 유지하는 것은 매우 어려운 작업이다. 효과적인 시그니쳐 생성을 위해 자동화 방법[8]이 제안되고 있지만, 실제 급변하는 네트워크 상황에서는 그 성능을 보장하기 어렵다.

위에서 제시한 시그니쳐 기반 분석 방법의 한계점을 보완하기 위해 본 논문에서는 플로우 상관관계 분석 방법을 제안한다. 상관관계 분석 방법은 시그니쳐 기반 분석 방법의 결과를 받아 분석된 플로우와 분석되지 않은 플로우간의 상관관계를 통해서 분석되지 않은 플로우를 분석하는 것이다. 총 4가지의 세부 방법으로 구성(서버-클라이언트, 발생 시간, 호스트-호스트, 통계 기반)되며 각 세부 방법은 하나의 분석기에서 통합되어 작동된다.

본 논문에서 제안하는 플로우 상관관계 기반 분석 방법론의 성능을 검증하기 위해 학내 망에 분석기를 설치하고 하루 동안 실험을 하였다. 그 결과, 시그니쳐 기반 분석성능(분석률)을 플로우 기준 약 10% 향상 시킬 수 있었다.

본 논문은 다음과 같이 구성되었다. Ⅱ장에서는 관련 연구를 살펴보며 Ⅲ장에서는 플로우 상관관계 기반 분석의 정의와 4가지 세부 방법에 대해 설명한다. 분석기의 시스템 구조와 자세한 구현 방법은 Ⅳ장에서 설명한다. Ⅴ장에서는 실험 및 결과를 Ⅵ장에서는 결론 및 향후 연구를 기술한다.

Ⅱ. 관련 연구

시그니쳐 기반 분석 방법에 비해 트래픽 상관관계 기반 분석 방법의 선행연구는 미흡한 상황이다. 본 장에서는 상관관계 기반 분석과 유사한 접근 방법을 가지는 몇 가지 선행 연구들을 소개한다.

BLNC[9]는 특정 응용을 사용 할 때 나타나는 호스트의 연결 상태를 그래프 형태로 나타내고 이를 패턴으로 정의하여 트래픽을 분석한다. 이러한 패턴은 세 가지 계층으로 나뉜다. (1)사회적 계층, (2) 기능적 계층, 그리고 (3)응용 계층이다. 이 논문의 특징은 페이로드와 포트번호와 같은 추가적인 트래픽 정보를 사용하지 않는다는 점이다. 즉, 응용 분석을 위해 호스트간 발생하는 트래픽의 연결 상태를 확인하고 이를 이용하여 분석한다. 따라서 기존 페이로드 시그니쳐 기반 분석 방법의 한계점인, 암호화 트래픽과 시스템 과부화의 문제점을 해결한다. 하지만 분석의 기준이 응용이 아니라 응용타입이라는 점과 호스트간의 연결 상태를 기반으로 하기 때문에 트래픽 수집 지점이 분석의 성능에 영향을 미친다는 단점을 가지고 있다.

[1]에서는 응용 트래픽의 패턴을 사용하여 응용을 분석한다. 이 방법은 분석 대상 네트워크에서 수집한 트래픽을 호스트간 세션과 포트의 개수를 기준으로 그룹화 하고 해당 그룹의 트래픽을 포트번호 정보를 통하여 분석하는 방법이다. 앞서 설명한 논문과 달리 응용의 타입이 아닌 응용 단위로 트래픽을 분석 할 수 있는 장점이 있지만, 그룹의 응용을 포트번호로 사용하여 결정하는 문제점이 있다. 최근 사용되고 있는 응

용들은 방화벽을 통과하기 위해 포트번호를 동적으로 할당한다. 따라서 포트번호를 사용한 응용 결정은 트래픽 분석의 정확도를 떨어뜨릴 수 있다.

본 논문에서는 선행 논문과 같이 트래픽의 상관관계를 이용하여 분석한다. 트래픽을 응용 기준으로 정확하게 분석하기 위해서 시그니쳐 기반 분석 방법의 결과를 이용한다. 따라서 정확도를 유지하면서 분석의 성능(분석률)을 극대화 할 수 있다.

Ⅲ. 플로우 상관관계 기반 분석

플로운 상관관계 기반 분석이란, 다른 분석기(시그니쳐 기반 분석기)에 의해 일부 분석된 결과를 입력받아 플로우에 나타나는 여러 정보 중 연관성을 가지고 있는 특정 속성 값을 기준으로 플로우들을 그룹화하여 분석된 플로우가 존재하는 경우 해당 그룹에 속한 전체 플로우를 분석된 플로우와 동일한 응용으로 분석하는 방법이다. 즉, 기존의 응용 시그니쳐 기반 분석기의 성능(분석률)을 향상시키기 위하여 분석된 플로우와 분석되지 않은 플로우 간의 상관관계를 이용하여 분석되지 않은 플로우를 분석하는 방법이다.

플로우 상관관계 기반 분석 방법의 가장 큰 장점은 선행되는 분석 방법(시그니쳐 기반 분석)이 분석하지 못하는 플로우를 분석된 플로우와 상관관계를 통하여 추가적으로 분석 할 수 있다는 점이다. 따라서 기존 시그니쳐 기반 분석기의 정확도를 유지하면서 분석률을 효과적으로 향상 시킬 수 있다. 특히, 트래픽이 암호화되어 있는 경우, 기존 시그니쳐 기반으로 분석하지 못하지만 본 방법론으로는 가능하다. 왜냐하면, 트래픽 암호화는 일부 트래픽 만에 적용되기 때문에 암호화되지 않은 트래픽을 시그니쳐 기반으로 분석한 경우, 이를 이용하여 암호화된 트래픽을 분석할 수 있다.

플로우 상관관계 방법의 고려사항은 다음과 같다. 첫째, 일정 양 이상의 분석된 결과가 입력 데이터로 필요하다. 본 분석 방법은 선행 분석 결과의 성능 향상(추가적인 분석)을 목적으로 하기 때문에 일정 양 이상의 분석된 결과가 필요하다. 상관관계를 이용하기 위해서는 연관성을 가지는 속성 값을 기준으로 그룹화하고 해당 그룹에 분석된 결과를 이용한다. 이때, 분석된 결과가 없을 경우에는 해당 그룹이 동일한 응용에서 발생된 플로우인 것은 알 수는 있지만, 어떠한 응용에서 발생되었는지는 알지 못한다. 둘째, 분석된 결과의 정확도가 보장되어야 한다. 입력 데이터가 부정확하면 본 방법론의 결과와 정확성을 보장 할 수 없을 뿐만 아니라 기존 분석 결과의 성능(정확도)을 저하 시킬 수 있다. 따라서 본 방법론의 입력 데이터(선행 분석기의 분석 결과)는 매우 정확하여야 한다.

본 논문에서는 플로우 상관관계를 이용하는 4가지 세부 분석 방법을 제안한다. 본 장에서는 각 방법들의 아이디어를 간단하게 살펴본 후, 다음 장에서 통합 분석기의 구조와 세부 방법의 자세한 구현 방법에 대해 알아본다. 본 논문에서 사용하는 플로우의 정의는 동일한 5-tuple(source address, source port, destination address, destination port, transport layer protocol)을 가지는 패킷과 그 역방향 패킷들의 집합이다[10].

표 1은 각 세부 방법에서 사용한 연관성을 가지는 속성 값을 정리한 것이다. 총 4가지 세부 방법들은 표 1에 제시한 속성값 기준으로 플로우를 그룹화 하고 각 그룹을 동일한 응용으로 분석한다.

표 1. 상관관계 방법에서 사용한 플로우 속성값

방법	속성값
서버-클라이언트 기반	address, port number, transport layer protocol(TCP/UDP)
발생 시간 기반	address, start time
호스트-호스트 기반	source address, destination address
통계 기반	address, number of packet, size of byte

3.1 서버-클라이언트 기반 상관관계

많은 사용자에게 인기 있는 응용과 웹사이트가 존재하기 때문에 해당 서버의 정보를 이용하면 분석률을 효과적으로 향상 시킬 수 있다. 서버-클라이언트 기반 상관관계 방법론은 다음과 같은 가정을 기반으로 한다.

■ 네트워크에서 발생하는 플로우 중, 특정 응용 서버 3-tuple(IP address, port number, transport layer protocol)을 공유하는 플로우들은 동일한 응용에 의해 발생되었을 가능성이 높다.

하나의 서버는 동일한 응용을 서비스하기 때문에 분석 결과를 서버 기준으로 그룹화하고 해당 그룹에 속한 모든 플로우를 동일한 응용으로 분석한다.

그림 1은 서버-클라이언트 기반 상관관계의 예시를 보여준다. 선행 시그니쳐 기반으로 특정 플로우(S1-C1)를 응용 "A"로 분석 하였을 경우, 서버 S1을 포함하는 다른 플로우(S1-C2, S1-C3)도 같은 응용으로 분석한다.

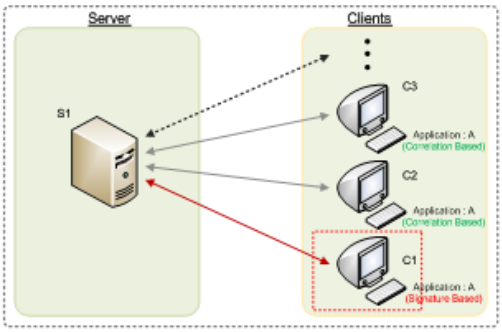

그림 1. 서버-클라이언트 기반 상관관계 예시

3.2 발생 시간 기반 상관관계

대부분 응용 트래픽은 사용자의 행동에 의해 트래픽을 발생된다. 따라서 특정 시점에 발생된 트래픽은 특정 응용에서 발생되었을 가능성이 높다. 이를 이용하면 분석률을 효과적으로 향상 시킬 수 있다. 발생 시간 기반 상관관계 방법론은 다음과 같은 가정을 기반으로 한다.

■ 특정 호스트에서 발생한 트래픽 중, 일정 기간 안에 발생하는 트래픽은 같은 응용일 가능성이 높다.

선행 시그니쳐 기반 방법으로 분석된 플로우를 호스트와 짧은 시간 간격으로 그룹화 한 후, 해당 그룹 중 분석된 플로우가 있을 경우 그 그룹에 속한 모든 플로우를 해당 응용으로 분석한다.

그림 2는 발생 시간 기반 상관관계의 예시를 보여준다. 선행 분석 결과를 각 호스트 별로 분류하고 시간에 대한 기준 값을 이용하여 플로우를 그룹화 한다. 이 때, 그룹화 기준은 플로우의 시작 시간으로 한다. 그룹화한 후, 각 그룹에 분석된 플로우가 포함되어 있으며, 해당 그룹의 모든 플로우를 분석된 플로우의 응용으로 분석한다. 그림에서와 같이 특정 호스트에서 발생한 트래픽을 일정 시간 기준으로 그룹화(Group A, B, C)하고 각 그룹에 선행 시그니쳐 방법으로 분석된 플로우가 존재하는지 확인한다. 만약 특정 응용으로 분석된 플로우가 존재한다면 해당 플로우의 분석 결과를 그룹 전체 플로우의 결과로 취한다.

이 방법은 그룹을 결정하는 일정 시간이 분석률과 정확도에 영향을 주기 때문에 분석 대상 네트워크에 맞는 적절한 시간 기준 값을 설정하기 위한 실험이 필요하다.

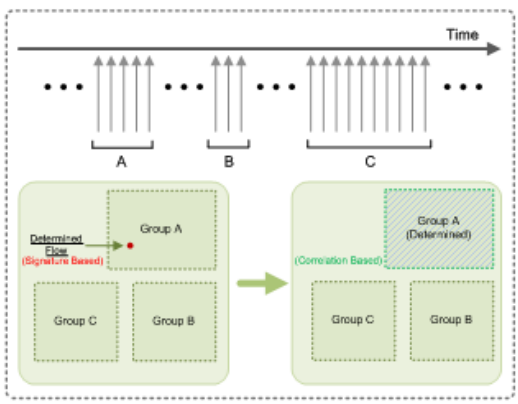

그림 2. 발생 시간 기반 상관관계 예시

3.3 호스트-호스트 기반 상관관계

데이터를 주고받는 통신일 경우 제어를 위한 통신과 실제 데이터를 전송하기 위한 통신이 서로 다른 포트에서 이루어지는 경우가 있다. 또한, 제어를 위한 트래픽과 달리, 데이터를 전송하기 위한 트래픽에서는 특정 용용의 시그니쳐를 추출하기 어려운 경우도 많다. 따라서 이러한 형태를 보이는 호스트들 사이의 트래픽을 호스트-호스트 상관관계 기반으로 분석한다. 호스트-호스트 기반 상관관계 방법론은 다음과 같은 가정을 기반으로 한다.

■ 같은 시간에 특정 두 호스트 간에 일어나는 통신은 하나의 응용에 의해 발생될 가능성이 높다.

특정 두 호스트 사이에서 발생한 플로우 중 일 부분만 시그니쳐 기반 분석 방법으로 분석이 되었다면, 분석되지 않은 플로우를 분석된 플로우의 응용으로 분석한다.

그림 3은 호스트-호스트 기반 상관관계의 예시를 보여준다. 선행 분석 결과를 각 호스트-호스트 별로 분류한다. 분류한 그룹 중 분석 결과가 존재하는 그룹은 해당 응용으로 분석한다. 그림 3과 같이 특정 두 호스트 간에 발생된 플로우 중 시그니쳐 기반으로 분석된 플로우가 존재하면, 분석되지 않은 플로우를 시그니쳐 기반 분석 결과를 이용한다.

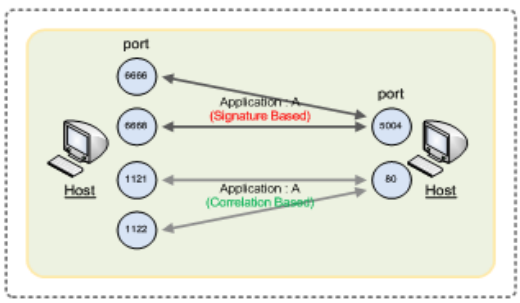

그림 3. 호스트-호스트 기반 상관관계 예시

3.4 통계 기반 상관관계

여러 호스트와 통신하는 P2P응용의 경우, 세션의 유지와 검색을 위하여 동일한 크기의 플로우를 발생한다. 이러한 특징을 이용하여, 특정 호스트에 한해, 같은 크기의 플로우들은 동일한 응용으로 분석한다. 통계 기반 상관관계 방법론은 다음과 같은 가정을 기반으로 한다.

■ 특정 호스트에서 발생한 플로우 중, 동일한 패킷 개수와 총 바이트 크기가 같은 플로우들과 동일한 응용일 가능성이 높다.

특정 호스트에서 발생한 플로우 중, 같은 크기를 가지는 플로우는 특정 응용의 동일한 기능을 수행하기 위해 발생된 것이다. 따라서 같은 크기를 가진 플로우들을 그룹화하고 그 중에서 분석된 플로우의 결과를 이용하여 그룹 전체 플로우를 분석한다.

그림 4는 통계 기반 상관관계의 예시를 보여준다. 선행 분석 결과를 호스트 별로 분류하고 분석된 결과와 같은 크기를 가지는 플로우를 선행 분석 결과와 동일한 응용으로 분석한다. 그림 4와 같이 시그니쳐 기반의 분석 방법으로 분석된 결과와 같은 개수의 패킷과 같은 크기의 바이트를 가지는 플로우가 존재 한다면 기존의 분석 결과를 이용한다.

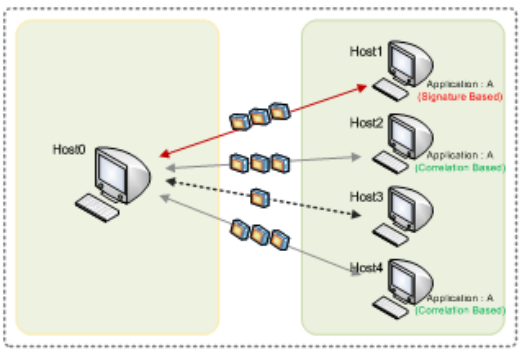

그림 4. 통계 기반 상관관계 예시

Ⅳ. 시스템 구조 및 구현

본 장에서는 앞장에서 제안한 플로우 상관관계 기반 트래픽 분석 세부 방법들을 적용한 트래픽 분석 시스템에 대해 기술한다.

4.1 상관관계 기반 분석 시스템

본 논문에서 제안하는 상관관계 기반 분석 시스템의 전체 구성도는 그림 5와 같다.

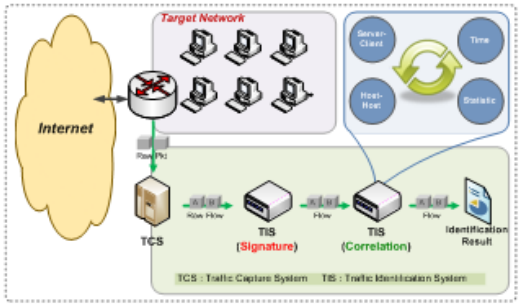

그림 5. 상관관계 기반 분석 시스템 구성도

분석 대상이 되는 네트워크와 인터넷 사이에 존재하는 백본 스위치에서 트래픽(패킷)을 수집한다. 수집된 트래픽은 플로우 형태로 조합하여 시그니쳐 기반 분석기로 입력된다. 본 시스템의 시그니쳐 기반 분석기는 본 연구진이 개발한 3가지(페이로드, 헤더, 통계) 시그니쳐를 기반으로 작동된다[3-5]. 시그니쳐 기반 분석기의 결과는 분석된 플로우와 분석되지 않은 플로우가 통합된 상태이며, 본 논문에서 제안하는 상관관계 기반 분석기에 입력된다. 입력된 플로우는 앞서 설명한 4가지 세부 분석 방법을 반복적으로 수행한다.

그림 6은 상관관계 기반 분석기의 내부 수행 과정을 순서도로 보여준다. 시그니쳐 기반 분석기의 분석 결과를 입력 받고 현재 분석 결과의 분석률을 pre 변

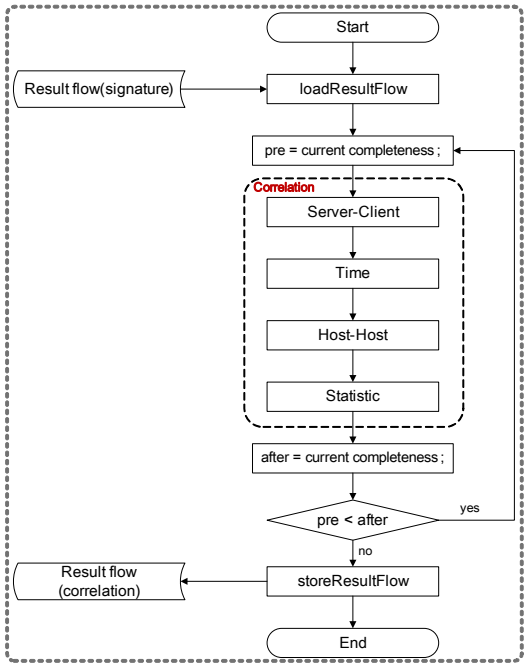

그림 6. 상관관계 기반 분석기의 내부수행 과정

수에 저장한다. 그리고 4가지 세부 방법을 순차적으로 수행한다. 모든 방법론이 수행된 후 현재 분석 결과의 분석률을 after 변수에 저장한다. 만약 after 변수에 저장된 값이 pre에 저장된 값보다 클 경우, 즉 추가적인 분석이 이루어진 경우에는 앞선 과정을 반복하게 된다. 상관관계를 통한 분석 방법은 기존 분석 결과를 이용하여 추가적인 분석을 가능하게 하기 때문에 각 분석 방법의 결과가 또 다른 추가적 분석을 가능하게 한다. 따라서 추가적인 분석이 더 이상 이루어지지 않을 때까지 4가지 세부 분석 방법을 반복한다.

4.2 개별 상관관계 분석 방법 구현

앞서 3장에서 설명한 4가지 세부 방법은 동일한 구현 방법을 가진다. 각 세부 방법에서 사용하는 속성값(표 1)들을 이용하여 선행 분석 결과를 그룹화하고 분석된 플로우를 통해 분석한다. 그림 7은 4가지 세부 분석 방법이 공통으로 사용하는 구현 방법을 보여준다. 본 논문에서 제안하는 4가지 방법론은 다른 분석기의 분석 결과를 입력 받아 플로우의 상관관계를 이용하여 추가적인 분석을 한다. 따라서 선행 분석된 결과를 로드하는 부분과 분석되지 못한 플로우를 분석하는 부분으로 나뉜다.

입력 데이터를 분석된 플로우와 그렇지 않은 플로우로 분류하고 각각 메모리에 로드한다. 로드 할 때에는 표 1에서 제시한 세부 방법 별 기준 속성 값을 사용하여 각 분석 방법에 맞게 그룹화 하여 로드한다. 예를 들어 서버-클라이언트 방법에서는 특정 호스트의 주소, 포트번호, 프로토콜 번호를 기준으로 그룹화

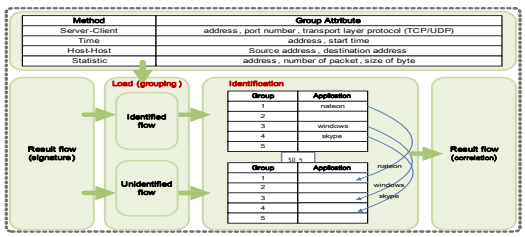

그림 7. 개별 상관관계 분석 방법 구현

하여 로드한다. 즉, 동일한 서버에서 발생된 플로우들을 하나의 그룹으로 만든다. 입력 데이터를 로드 한 후, 분석되지 않은 플로우로 구성된 그룹들 중 동일한 속성 값을 가진 분석된 그룹이 존재하면 해당 그룹의 응용을 분석 결과로 취한다. 즉, 분석된 그룹 결과를 통해 분석되지 그룹에 속한 트래픽을 분석한다. 분석된 결과 는 하나로 통합되어 저장 된다.

V. 실험 및 결과

본 장에서는 본 논문에서 제안하는 플로우 상관관계 기반 분석 방법의 성능을 확인하기 위해 실험한 내용을 기술한다. 우선 각 4가지 세부 방법의 개별 성능을 보이고 세부 방법을 통합한 상관관계 기반 분석기의 성능을 보인다.

평가 기준으로 분석률과 정확도를 사용하였다. 분석률은 전체 플로우 중 분석된 플로우의 비율을 의미하고 정확도는 분석된 결과가 얼마나 정확하게 분석되었는지를 의미한다.

$$\text{분석률} = \frac{Identified traffic}{Total traffic} \quad (1)$$

$$\text{정확도} = \frac{Correctly\ Identified traffic}{Total traffic} \quad (2)$$

표 2는 실험에서 사용한 트래픽 트레이스를 나타낸다. 하루 동안 학내 망에서 발생한 인터넷 응용(TCP/UDP) 트래픽을 수집하여 실험하였다.

표 3은 하루 동안 수집한 트래픽을 앞서 제안하는 상관관계 기반 분석 방법론으로 분석한 결과이다. 3장에서 설명한 총 4가지 방법의 성능을 확인하기 위하

표 2. 실험에 사용한 트래픽 트레이스 정보

Start time	Duration	Flow(K)	Packet(M)	Byte(G)
2010.04.01. 00:00	One day	69,076	1,929	1,549

표 3. 플로우 상관관계 기반 분석 결과

평가기준		시그니처	시그니처+ 서버-클라이언트		시그니처+ 발생시간		시그니처+ 호스트-호스트		시그니처+ 통계		시그니처+ {서버-클라이언트+발생시간+호스트-호스트+통계}	
분석률(%)	Flow	86.52	92.89	+6.27	89.29	+2.67	89.21	+2.59	87.22	+0.60	96.58	+9.96
	Byte	91.50	98.24	+6.74	91.63	+0.13	91.65	+0.15	91.51	+0.01	99.03	+7.53
정확도(%)	Flow	99.75	99.43	-0.32	99.43	-0.33	99.72	-0.03	99.71	-0.04	99.31	-0.44
	Byte	99.79	99.79	-0.01	99.78	-0.02	99.79	-0.01	99.79	-0.01	99.79	0.00

여 동일한 시그니쳐 기반 분석 결과[3-5]를 통합하여 입력 데이터로 사용하였다. 표 3의 각 열은 입력 데이터, 각 4가지 방법을 단독으로 적용한 결과, 그리고 입력 데이터에 4가지 방법을 통합하여 적용한 결과의 분석률, 정확도, 그리고 성능 향상률을 차례로 보여준다.

서버-클라이언트 기반 상관관계를 통해 기존 시그니쳐 기반 분석 결과를 분석 하였을 경우, 약 6%의 분석률 향상을 보였다. 서버-클라이언트 방법은 4가지 방법 중 가장 높은 성능 향상률을 보이는데 이것은 대다수의 사용자들이 특정 인기 있는 서버(포탈, 메일)를 많이 사용한다는 점을 반영한다. 즉, 사용자들이 많이 사용하는 응용 서버를 분석하면 복잡한 분석 방법을 적용하지 않고도 해당 응용 트래픽을 분석 할 수 있다.

시그니쳐 기반 분석 결과를 발생시간 기준으로 분석한 경우, 약 2.5%의 플로우를 더 분석 할 수 있었다. 3장에서 설명한 바와 같이 발생 시간 기반 상관관계 방법은 시간 기준 값이 필요하다. 정확한 시간 기준 값을 설정하기 위해 기준 값을 늘려가며 정확도를 측정하는 실험을 수행 하였다. 기준 값을 너무 짧게 설정하면 분석률 향상이 적고, 너무 길게 설정하면 서로 다른 응용 트래픽 그룹이 동일한 응용으로 잘못 분석되기 때문에 적절한 기준 값을 찾는 것이 중요하다. 실험결과 발생 시간을 기준으로 트래픽을 그룹화 할 때, 1msec를 기준 값으로 사용하는 것이 가장 최적의 결과를 보이는 것을 확인하였다. 하지만, 본 방법은 4가지 방법 중 가장 낮은 정확도를 보인다. 정확도 하락의 원인을 분석한 결과, 호스트에서 백그라운드로 동작하는 응용과 충돌하는 경우가 대부분이었다. 발생 시간 기반 분석 방법의 성능을 개선시키기 위해서는 적절한 시간 임계값을 찾는 방법과 호스트에서 발생하는 백그라운드 응용 트래픽을 분류하는 방법에 대한 연구가 필요하다.

호스트-호스트 기반 분석의 경우 약 3%의 분석률 향상을 보였고 통계 기반 분석은 약 0.5%의 분석률 향상을 보였다. 전체적으로 약간의 정확도 하락을 보이지만, 기존 시그니쳐 기반의 분석 결과의 성능(분석률)을 향상 시킨다.

표 3의 가장 오른쪽 열은 앞에서 설명한 4가지 방법을 모두 통합하여 구성한 플로우 상관관계 기반 분석기의 성능을 보여준다. 4가지 분석 방법은 더 이상 추가적인 분석이 없을 때까지 수행하였다. 결과적으로 플로우 기준 약 10%, 바이트 기준 약 7.5%의 분석률 향상과 약 0.5% 정확도 하락을 보였다. 기존의 시그니쳐 기반 분석 결과의 상관관계를 이용하여 분석되지 못한 플로우, 전체 플로우의 10%를 추가적으로 분석 할 수 있었다. 또한 수행 시간은 시그니쳐 기반 분석에 비해 약 0.05% 정도만 더 소요된 것을 확인하였다. 비록 약간의 정확도 하락이 존재하지만, 그 비율이 매우 낮고, 좀 더 세밀한 알고리즘 개선 작업이 이루어진다면 좋은 결과를 기대할 수 있다.

Ⅵ. 결론 및 향후 연구

인터넷 응용 트래픽 분석은 매우 중요하다. 기존의 시그니쳐 분석 방법은 널리 사용되고 있지만, 많은 한계점과 문제점을 가지고 있다.

본 논문에서는 시그니쳐 기반 분석기의 성능을 향상시키기 위해 플로우 상관관계 기반 분석 방법론을 제안하였다. 분석 방법론은 총 4가지의 세부 방법(서버-클라이언트, 발생 시간, 호스트-호스트, 통계 기반)으로 구성된다.

제안한 분석기를 실제 학내 네트워크에 설치하여 분석 한 결과 기존 시그니쳐 기반 분석 결과의 분석률을 약 10%이상 상승 시킬 수 있었다. 약간의 정확도 하락이 존재하므로 추가적인 알고리즘 개선 작업이 필요하다.

앞으로 각 세부 방법의 성능을 극대화 할 수 있는 방법을 연구하고, 플로우들의 상관관계를 통한 응용의 효과적인 분석 방법에 대해 연구할 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, 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〕 박준상, 박진완, 윤성호, 이현신, 김명섭, "응용 레벨 트래픽 분류를 위한 시그니쳐 생성 및 갱신 시스템 개발", *정보처리학회논문지* C 제17-C권 제1호, Feb., 2010, pp.99-108.

〔4〕 윤성호, 박진완, 박준상, 이상우, 김명섭, "고정 IP-port 기반 응용 레벨 인터넷 트래픽 분석에 관한 연구", *정보처리학회논문지* C 제17-C권 제2호, Apr., 2010, pp.205-214.

〔5〕 박진완, 윤성호, 박준상, 이상우, 김명섭, “통계 시그니처 기반의 응용 트래픽 분류”, *통신학회논문지* Vol.34 No.11, Nov., 2009, pp.1234-1244.
〔6〕 A. Dainotti, W. de Donato, and A. Pescap´e, “TIE: A Community-Oriented Traffic Classification Platform”, Proc. of the First International Workshop on Traffic Monitoring and Analysis, Berlin, Heidelberg, 2009.
〔7〕 M. Baldi, A. Baldini, N. Cascarano, and F. Risso, “Service-based traffic classification: Principles and validation”, Proc. of the IEEE 2009 Sarnoff Symposium, Princeton, NJ, USA, Mar., 2009.
〔8〕 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, Apri., 7-11, 2008, p.160-167.
〔9〕 Thomas Karagiannis, Konstantina Papagiannaki, Michalis Faloutsos, “BLINC: Multilevel Traffic Classification in the Dark”, Proc. of the SIGCOMM’05, Philadelpia, USA, 2005.
〔10〕 Cisco, NetFlow Services and Applications, White Paper, http://www.cisco.com/en/US/prod/collateral/iosswrel/ps6537/ps6555/ps6601/prod_white_paper0900aecd80406232.html.

윤 성 호 (Sung-Ho Yoon) 준회원

2009년 2월 고려대학교 컴퓨터정보학과 졸업
2011년 2월 고려대학교 컴퓨터정보학과 석사
2011년 3월~현재 고려대학교 컴퓨터정보학과 박사과정
<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

김 명 섭 (Myung-Sup Kim) 종신회원

1998년 2월 포항공과대학교 전자계산학과 졸업
2000년 2월 포항공과대학교 컴퓨터공학과 석사
2004년 2월 포항공과대학교 컴퓨터공학과 박사
2004년~2006년 Post-Doc., Dept. of ECE, Univ. of Torronto, Canada
2006년~현재 고려대학교 컴퓨터정보학과 부교수
<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석, 멀티미디어 네트워크