KNOM Review, Vol. 12, No. 2, Dec. 2009, pp. 12-21

페이로드 시그니쳐 기반 응용 레벨 트래픽 분류 시스템 성능 향상에 관한 연구

(Research on the Performance Improvement of Application-Level Traffic Classification System using Payload Signature)

박준상, 윤성호, 박진완, 이현신, 이상우, 김명섭

고려대학교 컴퓨터정보학과

{junsang_park, sungho_yoon, jinwan_park, hyunshin-lee, sangwoo_lee, tmskim}@korea.ac.kr

요 약

인터넷에 기반한 응용프로그램의 사용이 급격히 증가하면서 한정된 네트워크 자원을 효율적으로 사용하고, 사용자에게 안정적인 서비스를 제공하기 위한 많은 연구가 수행되고 있다. 이를 위해서는 다양한 종류의 응용 레벨 트래픽을 정확하게 분류할 수 있는 방법과 고속 링크에서 발생하는 대용량의 트래픽을 실시간으로 처리하는 방법에 연구가 선행되어야 한다. 트래픽의 분류을 위한 다양한 방법론이 존재하지만 Accuracy 와 Completeness 만을 고려했을 때 페이로드 시그니쳐 기반의 분석 방법은 가장 높은 성능을 보인다. 하지만 시그니쳐 추출 과정의 복잡성, 고속 링크에서의 처리 속도, 시그니쳐 유지 및 관리 문제는 페이로드 시그니쳐 기반 분석 방법의 실용성을 저하시키는 원인으로 작용한다. 본 논문에서는 선행 연구에서 제안한 페이로드 시그니쳐 자동 생성 시스템을 기반으로 추출한 시그니쳐의 Accuracy 와 Completeness 를 향상시킬 수 있는 페이로드 시그니쳐 관리 시스템을 제안하여 응용의 출현과 변화에 유연하게 대처할 수 있도록 한다. 또한 응용의 수와 대역폭의 증가로 인해 분류 시스템의 처리 속도가 지연되는 문제를 해결하기 위해 페이로드 시그니쳐의 계층 구조를 기반으로 트래픽을 분류하는 방법을 제안한다. 제안한 방법론은 다양한 응용 트래픽이 존재하는 고속 링크의 학내망에 실시간으로 적용하여 그 실효성을 증명한다.

Keywords: Payload Signature, Application-Level Traffic Classification

1. 서론

네트워크의 고속화와 더불어 전화망이나 전용망 기반의 음성 및 영상 서비스가 패킷 기반의 IP Network 에 통합되고, 다양한 서비스와 응용프로그램이 개발됨에 따라 기업이나 개인들은 인터넷으로 대표되는 네트워크에 대한 의존이 상당히 커져가고 있다.[7] 이와 같은 현실 속에서 네트워크의 효율적 운용과 관리를 위한 트래픽의 모니터링과 분석은 네트워크 사용현황 파악과 확장계획 수립 등의 전통적인 필요성 외에 다양한 분야에서 커져가고 있다.

응용 레벨 트래픽을 분석을 위한 다양한 방법이 존재하지만 Accuracy 와 Completeness 를 고려했을 때 페이로드 시그니쳐 기반의 분석 방법은 가장 높은 성능을 보인다.[3,4,5] 하지만 페이로드 시그니쳐를 수작업으로 추출하는 과정은 응용프로그램의 통합, 변경, 출현 등으로 인한 시그니쳐의 유지 및 관리에 대한 문제점으로 이어진다. 또한 현재의 페이로드 시그니쳐 기반 분석 방법은 고속 링크의 트래픽을 실시간으로 처리하는 과정에서 높은 부하를 발생시킨다. 응용의 개수와 네트워크 대역폭이 증가하는 추세를 고려했을 때 처리 속도의 향상은 반드시 해결해야 하는 과제이다.

따라서 본 논문에서는 선행 연구[1]에서 제안한 시그니쳐 자동 생성 시스템을 기반으로 시그니쳐를 생

이 논문은 2007 년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구(KRF-2007-331-D00387)와 2009 년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (2009-0090455)

성하고 에이전트를 통한 객관적인 검증 네트워크를 구축하여 생성 시스템의 실효성과 생성된 시그니쳐의 Accuracy 를 증명하고, 검증된 결과를 바탕으로 응용프로그램의 변화에 유연하게 대처할 수 있는 시그니쳐 관리 시스템을 제안한다. 또한 페이로드 시그니쳐를 응용 레벨 프로토콜 시그니쳐, 응용 프로그램 시그니쳐 두 단계 계층 구조로 표현하여, 분석 시스템의 시그니쳐 탐색 공간을 감소시켜 분류 시스템의 처리 속도를 향상 시킬 수 있는 방법을 제시한다.

본 논문의 구성은 다음과 같다. 본 장의 서론에 이어, 제 2 장에서는 기존 연구 방법의 문제점에 대 살펴보고, 3 장에서는 본 논문에서 제안하는 시그니쳐 관리 시스템을 설계하고 구현한다. 제 4 장에서는 학내 망에서 발생하는 전체 트래픽을 대상으로 제안하는 시스템을 구축하고 분석 결과를 통해 타당성을 증명한다. 마지막으로 5 장에서는 결론 및 향후 과제에 대해서 기술한다.

2. 관련 연구

표 1 은 기존의 다양한 분석 방법론에 대해서 응용 레벨 트래픽 분석의 성능을 평가하는 다양한 요소를 기준으로 비교 분석한 결과이다. 표 1 에서 보이는 각 분류 방법에 대한 성능 평가 결과는 분석 대상 네트워크에서 발생하는 모든 트래픽을 분류하고, 분류 시스템이 장기적으로 분석하는 것을 가정한다.

Method	Accuracy	Cost of Signature (Rule) Extract	Cost of Operation	Signature (Rule) Maintenance	Locality Dependence
	Completeness				Time Dependence
Well-Known Port	Low	Low	Low	Easy	Independent
	Low				dependent
Payload Signature	High	High	High	Difficult	Independent
	High				dependent
Machine Leaning	High	High	High	Difficult	dependent
	Mid				dependent
Flow Correlation	High	High	High	Difficult	dependent
	Low				dependent
Application Behavior	High	High	High	Difficult	Independent
	Low				dependent

표 1. 응용 레벨 트래픽 분류 방법론 성능 비교

표 1 과 같이 페이로드 시그니쳐 기반 분석 방법은 분류의 Accuracy 와 Completeness 측면에서 다른 분석 방법론에 비해 상대적으로 가장 높은 성능을 나타낸다. 하지만 시그니쳐 추출 과정의 복잡성, 분류 시스템의 부하, 시그니쳐의 지속적인 유지가 어려운 문제점을 보인다. 이러한 문제점을 보완하기 위해 표 2 와 같이 기존의 다양한 연구에서 해결책을 제안하고 있지만 아직까지 다양한 문제점이 나타나고 있다.

해결 과제	기존 연구	기존 연구의 문제점
시그니쳐 추출 과정 복잡	Park et al. [4] "Towards Automated Application Signature Generation for Traffic Identification"	Common Keyword 추출 Main Traffic으로 제한
	Haffner et al. [5] ”ACAS: automated construction of application signatures”	수동적인 데이터 수집 Ground Truth 보장하지 못함. 고정된 위치에서 추출
	Xiao et al. [6] “ASG - Automated signature generation for worm-like P2P traffic patterns”	
시그니쳐 유지	N/A	
분류 시스템 처리속도 지연	Sen et al. [3] “Accurate, scalable in-network identification of p2p traffic using application signatures”	기존에 존재하는 스트링 매칭 알고리즘 성능 비교.
	Risso et al. [8] "Lightweight, Payload-Based Traffic Classification An Experimental Evaluation"	
	Yongmin Choi. [9] “On the Accuracy of Signature-based Traffic Identification Technique in IP Networks"	

표 2. 페이로드 시그니쳐 기반 분석 방법의 문제점에 대한 기존 연구 결과

수작업으로 시그니쳐를 추출하는 과정의 복잡성을 해결하기 위해 페이로드 시그니쳐의 자동 생성 방법에 관한 연구[5, 6]가 제안되었다. 자동화된 추출 방법으로 페이로드 데이터로부터 동일한 스트링을 추출하

는 알고리즘을 제시하고 있지만 수동적인 데이터의 수집 방법을 사용한다. 시그니쳐는 Ground Truth 의 정확성에 따라 성능이 좌우되기 때문에 정확한 데이터 수집 방법이 전제되어야 한다. 또한 제안된 방법은 트래픽을 응용 레벨 프로토콜을 기준으로 분류하는 것을 목적으로 한다. 하지만 응용 레벨 프로토콜에 의한 분류 기준은 HTTP Tunneling 과 같이 트래픽을 은닉하여 정보를 전달하는 응용프로그램의 트래픽의 분석이 어렵고, 응용프로그램의 통합화로 하나의 응용프로그램이 다양한 종류의 프로토콜에 기반하여 서비스되기 때문에 프로토콜 기준의 분류는 네트워크 관리자에게 효과적인 정보를 제공하기 못한다.

[4]는 LCS 알고리즘에 기반한 시그니쳐 자동 생성 시스템을 제시하고 있다. 하지만 LCS 의 입력 데이터에 대한 제약사항으로 패킷 크기만을 비교하여 적용하고 있다. 이러한 방법은 패킷 크기에 대한 임계값 설정에 어려움이 발생하고, 다른 기능을 수행하는 트래픽이 동일한 패킷 크기를 갖는 경우 시그니쳐 생성이 불가하거나, 잘 못된 시그니쳐의 추출 가능성이 높다. 또한 시그니쳐로서 추출하는 정보가 스트링으로 제한되어 있다. 페이로드 시그니쳐의 Accuracy 를 향상 시키고, 분류 시스템의 부하를 줄이기 위해서는 패킷의 순서, 페이로드 offset 정보에 대한 추가적인 정보의 추출이 요구된다.

응용 프로그램은 다양한 서비스를 통합하여 사용자에게 제공한다. 따라서 응용의 변화와 업데이트가 잦고 새로운 응용이 많이 등장하고 있다. 이는 응용 프로그램 시그니쳐의 Accuracy 를 지속적으로 유지하기 어렵게 만드는 요소로 작용한다. 하지만 페이로드 시그니쳐 기반의 기존 연구에서는 대부분 시그니쳐의 추출에 초점을 두고, 시그니쳐의 관리를 위한 방법을 제시하지 못하고 있다. 따라서 본 논문에서는 응용 프로그램 시그니쳐를 갱신하고, 새로운 응용의 시그니쳐를 추출하기 위한 방법을 제시한다.

페이로드 기반 분석 방법은 Accuracy 가 높지만 Well-Known Port 기반 분석 방법, 고정 IP, Port 기반 분석 방법에 비해 분류 시스템의 처리 속도가 느려서 트래픽에 컨트롤을 위한 목적으로 사용하기에 부적합한 방법이다. 과거에는 사용하는 응용의 개수가 적고 트래픽의 발생량이 적기 때문에 분류 시스템의 처리 속도가 문제가 되지 않았지만 현재는 응용의 개수가 증가하고 대용량의 트래픽을 발생시키는 응용의 사용이 증가하면서 분류 시스템의 처리 속도는 페이로드 기반 분석 방법에서 해결해야 하는 과제이다. 이러한 문제점을 해결하기 위한 연구[3,8,9]가 수행되었지만 이러한 방법들은 기존에 존재하는 스트링 매칭 알고리즘 Library 에 대한 비교 분석에 그치고 있다. 본 논문에서는 페이로드 시그니쳐의 구조를 이용하여 시그니쳐의 탐색 공간을 줄임으로서 분류 시스템의 처리 속도를 향상 시키는 방법을 제안한다.

우리는 선행 연구[1]에서 수작업으로 시그니쳐를 생성하는 문제를 해결하기 위해 시그니쳐 자동 생성 시스템을 제안하였다. 시그니쳐 생성 시스템을 이용하여 학내 망에서 발생하는 응용 중 129 개의 응용에 대해 838 개의 시그니쳐를 추출하였다. 그림 1 은 시그니쳐 생성 시스템을 통해 추출한 시그니쳐를 2009 년 08 월 27 일 00:00 부터 2009 년 09 월 02 일 23:59 까지 학내 망에서 발생하는 연속적인 트래픽에 적용하여 Accuracy 와 전체 트래픽에 대한 Completeness 를 측정한 결과이다.

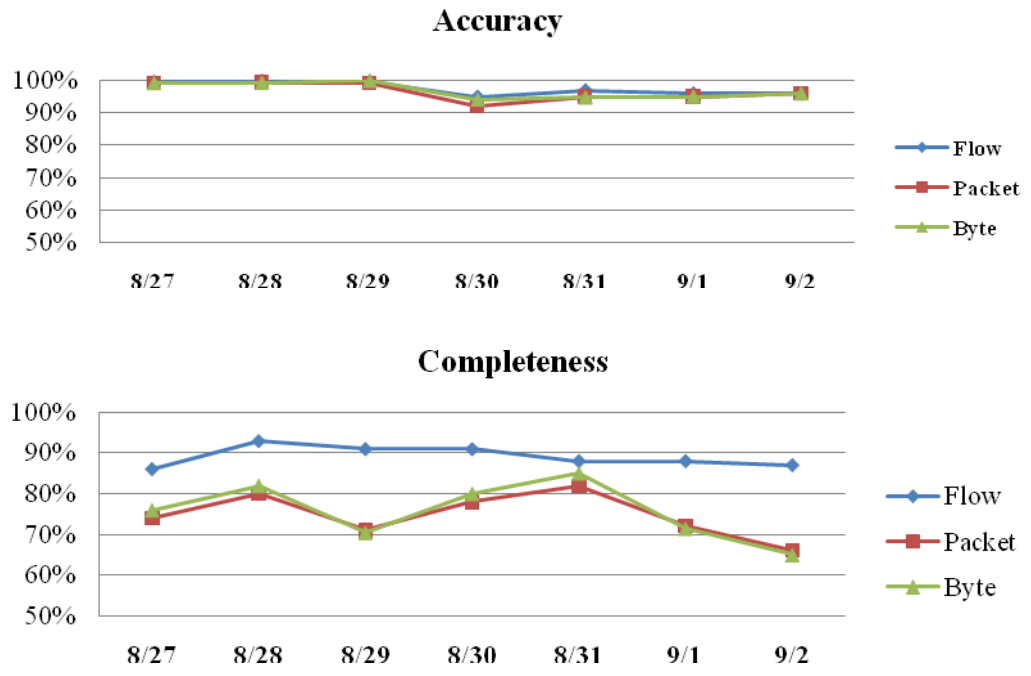


그림 1. 시그니쳐 생성 시스템에 기반한 트래픽 분류 결과

Accuracy 는 분류된 트래픽에 대한 정확도를 나타낸다. 측정 결과 92%이상의 높은 Accuracy 을 보이는 것을 것 수 있다. 하지만 Completeness 측정 결과 안정적인 분석률을 나타내지 않으며, 09 월 01 일부터 Packet/Byte 의 Completeness 가 70%이하로 감소되는 것을 알 수 있다. Accuracy 를 고려했을 때 대부분 분류 범위(Coverage) 내에 존재하지 않은 응용의 트래픽으로 분석되었다. 09 월 01 일 이후 Completeness 의 감소 원인은 캠퍼스 네트워크의 특성상 개강 직후 트래픽의 양이 증가하고 새로운 응용의 발생 빈도가 많기 때문이다. 이와 같이 분석 대상 네트워크에서 발생할 수 있는 응용은 예측하기 어렵기 때문에 지속적으로 Completeness 를 유지할 수 없다. 따라서 시그니쳐 관리 시스템은 응용 레벨 트래픽 분석 방법에 있어 반드시 요구된다.

3. 시스템 설계 및 구현

본 장에서는 논문에서 제안하는 시그니쳐 관리 시스템에 대해 기술하고, 시그니쳐 기반 분석 시스템의 성능 향상 방법에 대해 기술한다.

3.1. 시그니쳐 관리 시스템

본 절에서는 응용의 변화에 유연하게 대처할 수 있는 페이로드 시그니쳐 관리 시스템의 구성에 대해 살펴보고 관리 시스템의 핵심이 되는 피드백 시스템의 구성과 제공되는 정보에 대해 기술한다.

3.1.1 시그니쳐 관리 시스템의 구성

시그니쳐 관리 시스템은 시그니쳐의 검증 결과를 바탕으로 응용의 변화를 인지하고 관리자에게 시그니쳐를 갱신할 수 있는 트래픽과 정보를 제공한다. 시그니쳐 관리 시스템은 시그니쳐 추출, 트래픽 분류, 검증 데이터 수집, 피드백 시스템으로 구성된다. 그림 2 는 관리 시스템의 전체 구조를 나타내고 있다. 추출 시스템은 선행 연구[1]에서 제안한 시그니쳐 자동 생성 시스템으로 구성된다. 트래픽 분류 시스템은 페이로드 시그니쳐를 기반으로 학내 망에서 발생하는 모든 트래픽을 분류한다. 피드백 시스템은 분류 시스템의 분류 결과와 TMA(Traffic Measurement Agent)에 기반한 Ground Truth 데이터를 바탕으로 객관적으로 검증하고, 오 분류되거나 미 분류된 트래픽을 시그니쳐 생성 시스템의 입력 데이터로 제공한다.[1]

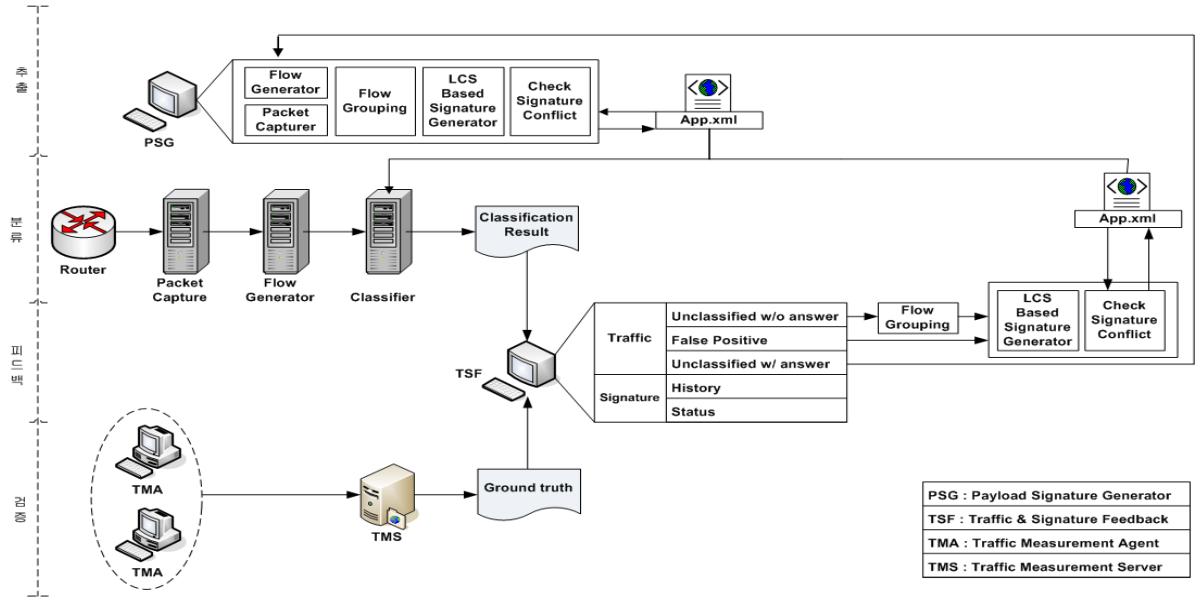


그림 2. 시그니쳐 관리 시스템 구조

3.1.2. 피드백 시스템의 구성 및 적용 범위

시그니쳐 관리 시스템은 핵심적인 부분은 분류된 트래픽에 대한 객관적인 검증을 통해 오 분류되거나 미 분류된 트래픽에 대한 정보를 제공하는 피드백 시스템이다. 그림 3 은 피드백 시스템의 시그니쳐의 관

리 목적에 따른 적용 범위를 표현하고 있다. 분류 대상 전체 트래픽을 분류한 트래픽, 분류 가능한 트래픽, Ground Truth 를 갖는 트래픽으로 구분하고 각 범위 내에서 시그니쳐 추출, 갱신, 삭제를 위한 정보를 제공한다.

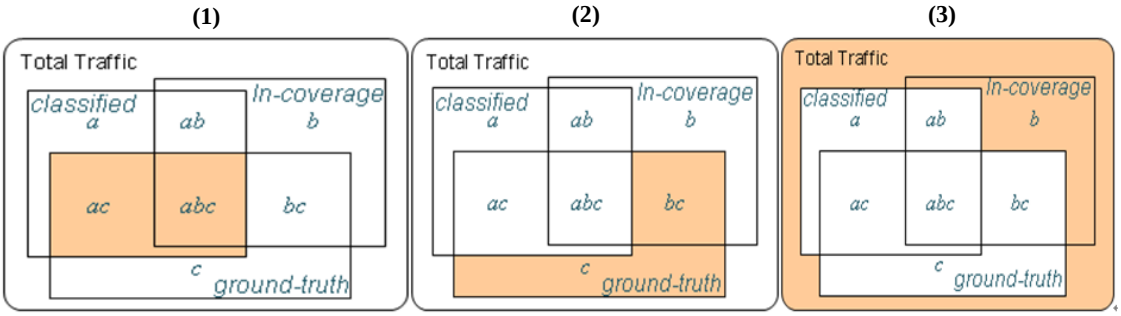


그림 3. 피드백 시스템의 목적에 따른 트래픽 범위

그림 3 의 (1)의 분류 결과를 바탕으로 잘 못 분석한 트래픽을 관리자에게 제공한다. 이러한 트래픽을 통해 관리자는 응용의 시그니쳐를 수정하거나 새로운 시그니쳐를 추가할 수 있다. (2)의 영역에 포함되는 트래픽을 제공하여 분석 범위에 포함되지 않은 응용을 찾고 관리자에게 해당 응용의 이름과 트래픽 정보를 제공한다. (3)의 영역에 포함되는 트래픽은 알 수 없는 응용에 의해 발생된 트래픽이거나, 분류 범위에 포함되어 있지만 응용의 변화로 인해 분류하지 못하는 트래픽이다. 이러한 트래픽은 관리자에게 제공되고 페이로드 정보를 기반으로 새로운 응용을 찾아내거나 응용의 시그니쳐를 갱신하는 목적으로 사용된다.

표 4 는 피드백 시스템의 구성과 역할에 대해 설명하고 있다. 그림 3 에 나타내는 관리 시스템의 목적에 따른 트래픽의 범위를 함께 나타낸다.

구성	구분	역할	범위
Signature Feedback System	Signature History	시그니쳐의 충돌 관계 확인	(1)
	Signature Status	시그니쳐의 유효성 판단 기준	(1)
Traffic Feedback System	False Positive	오 분류된 트래픽 시그니쳐 수정	(1)
	Unclassified w/ Answer	새로운 응용의 출현 보고	(2)
	Unclassified w/o Answer	Unknown Type의 시그니쳐로 정의	(3)

표 4. 피드백 시스템의 구성과 역할

피드백 시스템은 시그니쳐 피드백 시스템과 트래픽 피드백 시스템으로 구분된다. 시그니쳐 피드백 시스템은 2 가지 역할을 수행한다. 그림 3 의 (1)의 범위의 포함되는 플로우를 분류한 시그니쳐의 분류 로그 정보를 제공하는 기능과 (1)의 영역의 트래픽을 정확하게 분류한 각 시그니쳐의 분석량에 대한 정보를 제공하요 각 시그니쳐의 유효성을 판단할 수 있는 근거로 사용된다. 피드백 시스템의 트래픽 피드백 시스템은 그림 3 과 같은 범위의 트래픽에 대해서 오분류되거나 미분류되어진 플로우를 수집하여 관리자에게 제공한다. 관리자는 트래픽을 플로우 그룹핑 단계, LCS 기반의 시그니쳐 추출 단계, 시그니쳐 유효성 평가 단계를 거쳐 응용프로그램의 시그니쳐를 갱신한다. TMA 를 기반으로 수집된 Groud Truth 정보는 새로운 응용의 출현에 대한 정보를 제공하고 리포팅을 통해 시그니쳐 추출 대상 응용프로그램을 선정하고 시그니쳐 생성 시스템을 기반으로 시그니쳐 생성하고 등록하게 된다.

3.2. 분류 시스템의 처리 속도 향상 방법

본 절에서는 페이로드 시그니쳐를 응용 레벨 프로토콜 시그니쳐, 응용 프로그램 시그니쳐의 구조를 두 단계 계층으로 표현하여 탐색 공간과 비교 대상 시그니쳐의 수를 줄여 분류 시스템의 처리 속도를 향상 시킬 수 있는 방법을 제시한다.

3.2.1 페이로드 시그니쳐 계층 구조

페이로드 시그니쳐 생성 시스템 및 관리 시스템을 기반으로 학내 망에서 사용되는 139 개의 응용프로그램에 대해 845 개의 시그니쳐를 생성하였다. 모든 시그니쳐는 트래픽의 분류 기준인 응용 프로그램 단위로 생성되었다. 응용 프로그램 단위의 시그니쳐는 공통의 응용 레벨 프로토콜 시그니쳐를 포함하는 형태를 보

인다. 따라서 그림 4 와 같이 응용 레벨 프로토콜 시그니쳐, 응용 프로그램 시그니쳐의 2 단계 계층 구조로 표현 할 수 있다. 이와 같은 계층 구조는 분류 시스템의 탐색 공간을 줄이고, 응용 레벨 트래픽의 분류 기준을 응용 레벨 프로토콜과 응용 프로그램으로 분석 가능한 유연성을 제공할 수 있다.

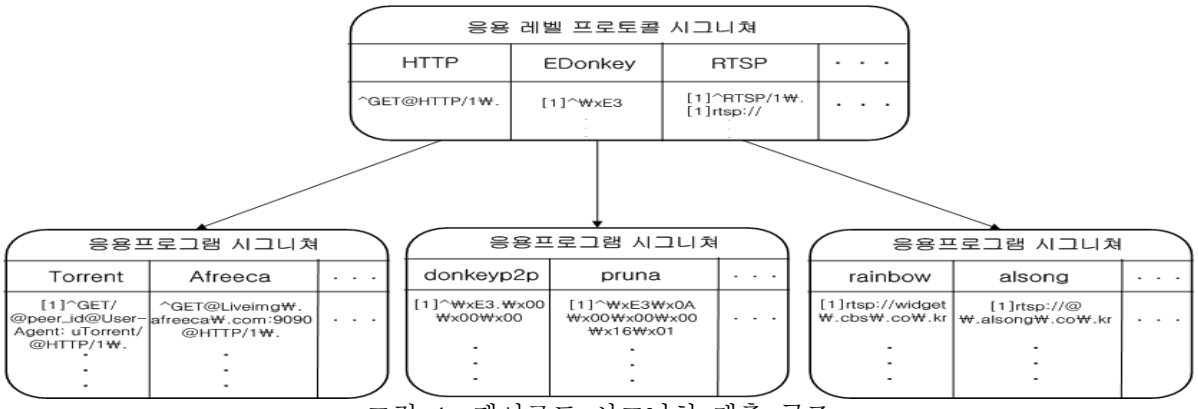


그림 4. 페이로드 시그니쳐 계층 구조

3.2.2 트래픽 분류 방법

트래픽 분류 시스템의 입력 데이터는 패킷 데이터를 포함하는 양방향의 플로우와 계층적 구조로 표현된 페이로드 시그니쳐이다. 분류 시스템은 2 단계 분류 과정을 통해서 최종적으로 해당 응용의 플로우를 분석하게 된다. 그림 5 는 트래픽 분류 시스템의 구성과 데이터의 처리 과정을 도식화 한 것이다.

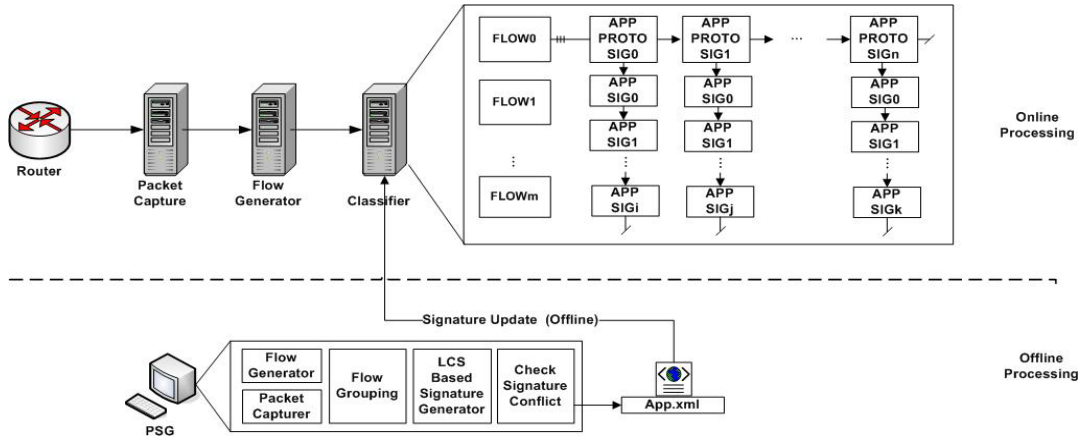


그림 5. 분류 시스템 구조

페이로드 시그니쳐 기반 분류 시스템의 입력 플로우는 1 차적으로 응용 프로토콜 레벨 시스니쳐를 우선적으로 비교한 후, 매칭되는 시그니쳐가 존재하는 경우 해당 응용 레벨 프로토콜 시그니쳐에 포함되는 응용프로그램 시그니쳐만을 비교하여 최종적으로 응용프로그램 단위로 분류하게 된다. 기존의 방법은 분류 시스템의 입력 플로우에 대해 모든 시그니쳐를 비교하여 플로우를 분류하기 때문에 분류 시스템의 부하가 발생하고 분석 시간의 지연이 발생하였다. 하지만 제안한 방법은 모든 시그니쳐를 비교하지 않기 때문에 시그니쳐의 탐색 공간을 줄일 수 있어 처리 속도를 향상 시킬 수 있다.

4. 실험 및 결과 분석

본 장에서는 3 장에서 기술한 시그니쳐 관리 시스템을 학내 망에 적용하여 Accuracy 와 Completeness 를 통해 그 타당성을 증명한다. 또한 분류 시스템의 처리 속도에 대한 결과를 기존의 선형적인 분류 방법과 제안하는 계층적 구조의 분류 방법을 비교하여 성능을 평가한다.

4.1. 트래픽 분류 결과

Coverage, Accuracy, Completeness 는 시그니쳐의 실효성 및 정확성을 평가하기 위한 Metric 으로 사용된다. 시그니쳐의 검증의 결과를 평가하기 위한 트래픽 데이터는 학내 망에서 발생하는 인터넷 트래픽이며, 2009 년 12 월 06 일 00:00 분부터 2009 년 12 월 12 일 11:59 분까지 1 주일의 연속적인 트래픽으로 평가되었다. 학내 망에는 P2P 파일 공유 프로그램, P2P 메신져, 인터넷 디스크 등 일 단위 200 개 이상의 다양한 응용프로그램이 분석되는 것을 알 수 있었다.

4.1.1. Coverage

시그니쳐 생성 시스템을 이용하여 학내 망에서 발생하는 응용프로그램들을 대상으로 표 6 과 같은 분류 범위를 갖는 시그니쳐를 추출하였다. 2009 년 09 월 02 일 이후 지속적으로 관리 시스템을 적용한 결과 46 개의 새로운 응용이 추가되었다. 이 중 11 개의 응용프로그램이 새로운 응용으로 등록되었고 35 개는 Unknown Type 의 응용으로 추출되어 98 개의 시그니쳐가 추가되었다. 새로운 응용으로 등록된 11 개의 응용프로그램은 그 발생 빈도는 적었지만 대량의 트래픽을 발생하는 P2P File Sharing, Web Disk 가 가장 많은 비율을 보였다.

분류 단위	분류 대상 개수	
	09.02 00:00	12.06 00:00
응용프로그램	129	140
Unknown Type 응용	0	35
프로토콜	13	13
프로세스	204	265
시그니쳐	838	935

표 6. 관리 시스템 적용 전·후 Coverage 변화

4.1.2. Completeness

그림 6 은 실험 기간 동안 학내 망 전체에서 발생된 트래픽 중 분류된 트래픽의 양을 나타내고 있다. 그림 1 과 같이 관리 시스템을 적용하기 전 09 월 02 일 Completeness 는 Byte/Packet 기준으로 70% 이하의 Completeness 를 보이고 불안정한 형태로 나타났다. 관리 시스템의 적용 후 Byte/Packet 기준으로 80% 이상의 Completeness 를 지속적으로 유지할 수 있는 것을 알 수 있다.

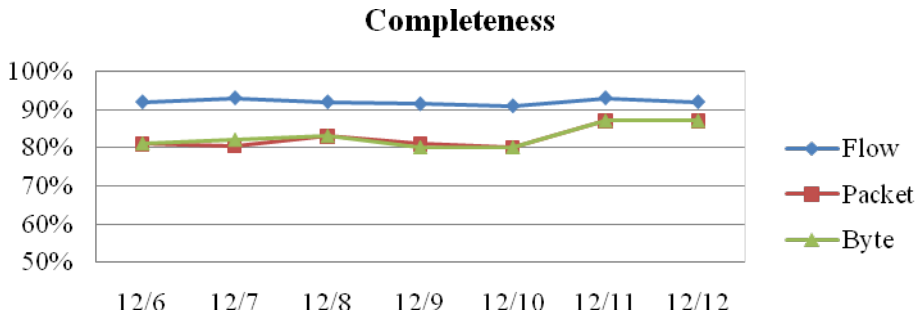


그림 6. Completeness 결과

Completeness 를 감소시키는 원인은 두 가지 경우로 분석된다. 첫째, 페이로드의 암호화로 시그니쳐를 추출하기 어려운 경우이다. 현재까지 분석된 응용프로그램 중 Skype, BitTorrent 등 시그니쳐가 생성되지 않거나 일부 트래픽에 대해서만 시그니쳐 분석이 가능한 응용프로그램은 Completeness 를 감소시킨다. 둘째, 분석되지 않은 응용 프로그램 즉 시그니쳐 기반 분류 시스템에 등록되지 않은 응용프로그램에 의해서 Completeness 가 감소된다. 시그니쳐 기반의 분류 방법은 응용프로그램을 분류하기 위해서 선행적으로 분류 대상 응용프로그램의 분석을 통해 해당 시그니쳐를 추출하는 과정이 요구된다. 응용 프로그램을 알 수 있다면 시그니쳐 생성 시스템에 기반하여 시그니쳐의 추출이 가능하며 분류할 수 있다. 하지만 응용프로그램을 알 수 없는 경우 시그니쳐의 추출이 불가능하기 때문에 Completeness 를 감소시키는 가장 큰 원인이 된다. 관리 시스템은 응용을 알 수 없는 경우 Ground Truth 를 기반으로 응용에 대한 정보를 제공하거나, Ground Truth 에 나타나지 않는 응용에 대해서는 Unknown Type 의 응용으로 분류할 수 있는 정보

를 제공할 수 있기 때문에 Completeness 향상을 위해 반드시 요구된다.

4.1.3. Accuracy

그림 7 은 시그니쳐의 정확성을 평가하는 Accuracy 에 대한 측정 결과를 나타낸다.

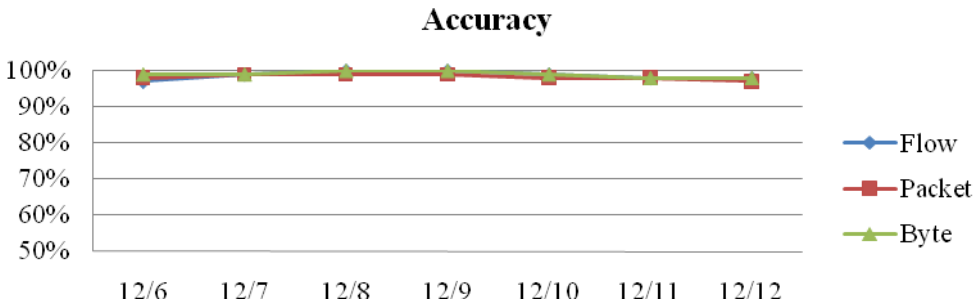


그림 7. Accuracy 결과

Accuracy 의 측정 결과를 통해서 시그니쳐 관리 시스템의 중요성을 확인할 수 있다. 시그니쳐 생성 시스템에 의해 추출한 시그니쳐만을 기반으로 학내망 전체 트래픽을 분류한 후 관리 시스템을 적용하여 오분류되거나 미분류된 플로우를 대상으로 관리시스템을 통해 시그니쳐를 갱신하였다. 그 결과 시그니쳐의 정확도가 향상되었고 99%이상의 높은 Accuracy 를 유지하고 있는 것을 확인할 수 있었다. Accuracy 를 감소시키는 원인은 세 가지로 분석되었다. 첫째, Internet Explorer 와의 충돌로 인해 오 분류되는 경우가 가장 높은 비율을 점유했다. 이는 특정 응용프로그램 내에 Internet Explorer 의 트래픽이 임베딩된 형태로 발생되어 동일한 시그니쳐의 형태를 갖기 때문이다. 응용프로그램이 Internet Explorer 에 대한 의존도가 커져가고 있는 추세를 고려했을 때 시그니쳐 기반의 분석 방법에 대한 Accuracy 는 점차 감소할 것으로 판단된다. 둘째, 패킷 내에 페이로드가 존재하지 않는 트래픽으로 인해 미분류되는 경우가 발생했다. 페이로드가 존재하지 않는 트래픽은 해당 네트워크에 크게 영향을 미치지 않는다는 이유로 기존의 연구에서는 트래픽의 분류대상에서 제외 시켰지만, 학내망의 트래픽에 대해 TCP 의 연결 설정이 맺어진 후 연결 종료가 이루어진 플로우를 대상으로 그 비율을 조사한 결과 전체 플로우의 5.47%의 무시할 수 없는 수준의 비율을 보였다. 이러한 트래픽의 분류를 위해 헤더 시그니쳐 기반의 분류 방법을 적용하였지만 분류가 불가능하였다. 셋째, 패킷의 페이로드가 존재하지만 페이로드의 암호화로 인해 스트링의 특정한 패턴을 찾을 수 없는 플로우는 시그니쳐로서 분류 되지 못하고 Accuracy 를 감소시켰다.

4.2. 시그니쳐 기반 분석 시스템 처리 속도 향상

트래픽 분류 시스템의 성능 분석을 위해 학내망에서 발생하는 모든 트래픽을 평면적 구조의 처리 방식과 계층적 구조의 처리 방식을 적용하여 처리 시간을 비교하였다. 두 방식은 동일한 트래픽 트래이스를 대상으로 적용되었고, 표 6 의 9 월 02 일과 같은 분류 범위를 갖는다.

그림 8 은 트래픽 양이 급격히 증가하는 오후 12:00 부터 12:59 까지의 연속적인 트래이스에 대해 1 분 단위 데이터에 대한 처리 시간을 나타내고 있다. 3,300,090 개의 플로우와 약 66GB 의 트래픽을 분류한 결과이다. 분석된 응용의 개수는 113 개로 나타났으며 다양한 종류의 Web Disk, P2P 트래픽이 존재하였다.

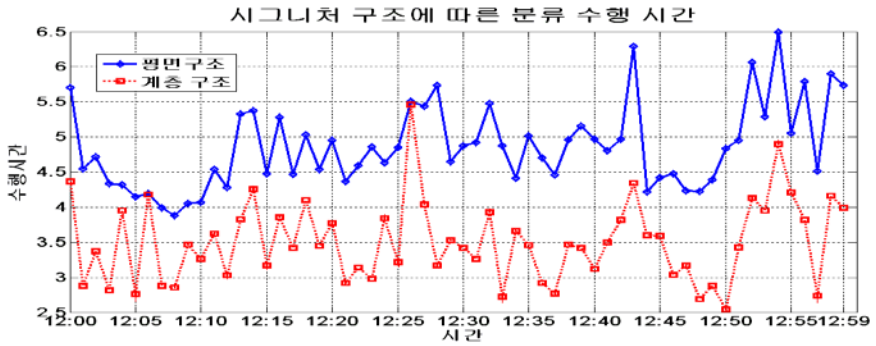


그림 8. 분류 시스템 처리 시간

두 가지 방법 모두 트래픽의 양과 응용 프로그램의 수가 증가 할수록 처리시간이 길어지는 것을 알 수 있다. 하지만 평면적인 분석 방법에 비해 계층적 구조의 분석 방법이 평균 10 초 정도 빠른 처리를 하는 것을 알 수 있다. 이는 분류 시스템에서 처리해야 하는 시그니쳐의 비교 횟수가 감소했기 때문이다. 제안한 방법은 시그니쳐의 개수와 트래픽의 양이 증가할수록 순차적인 분석 방법에 비해 상대적으로 높은 성능을 보인다.

5. 결론 및 향후 과제

페이로드 시그니쳐 기반 분석 방법은 높은 Accuracy 와 Completeness 는 높지만 시그니쳐 생성 후 관리 문제, 분류 시스템의 처리 속도 문제로 그 실효성이 저하되는 문제점을 갖는다. 따라서 본 논문에서는 응용 프로그램에 대한 사용자의 기호 변화, 응용의 갱신, 출현에 따른 시그니쳐의 Accuracy 와 Completeness 가 감소되는 문제에 대처하기 위해 객관적인 검증에 기반한 시그니쳐 관리 시스템을 구축하였다. 생성된 페이로드 시그니쳐는 응용 레벨 프로토콜, 응용 프로그램의 2 단계 계층 구조로 표현되어 실시간 분류가 가능하도록 하였다. 제안한 시스템은 학내 망에서 발생하는 전체 트래픽을 실시간으로 분석 가능하였고, 99% 이상의 Accuracy 와 80% 이상의 Completeness 를 지속적으로 유지할 수 있었다.

페이로드 기반 분석 방법의 Completeness 를 향상 시키기 위해서는 페이로드가 존재하지 않는 플로우나, 암호화된 트래픽에 대한 분류가 반드시 요구된다. 이러한 트래픽은 플로우의 다양한 상관 관계를 기반으로 분류가 가능하다. 응용 프로그램 트로토콜을 의미적으로 접근하여 플로우 사이의 상관 관계를 정의하여 페이로드가 존재하지 않거나, 암호화된 트래픽에 대한 연구가 요구된다.

응용프로그램의 변화와 출현에 대해 네트워크 관리자가 수동적으로 인식하고 대처하기에는 많은 인력과 시간이 요구된다. 이러한 문제점을 해결하기 위해 검증 네트워크에 데이터 수집, 시그니쳐 추출, 시그니쳐 갱신 시스템을 통합하고 시그니쳐를 데이터베이스화 하여 주기적으로 시그니쳐의 갱신이 이루어지는 시스템을 구축할 계획이다.

6. 참고 문헌

[1] 박준상, 박진완, 윤성호, 오영석, 김명섭, "응용 레벨 트래픽 분류를 위한 시그니쳐 생성 시스템 및 검증 네트워크의 개발", 2009 년 제 31 회 정보처리학회 춘계학술발표대회 (KIPS), 부산, 한화리조트, Apr. 23-24, 2009, 제 16 권 제 1 호, pp. 1288-1291

[2] Myung-Sup Kim, Young J. Won, and James Won-Ki Hong, "Application-Level Traffic Monitoring and an Analysis on IP Networks," ETRI Journal, Vol.27, No.1, Feb. 2005, pp.22-42.

[3] Subhabrata Sen , Oliver Spatscheck , Dongmei Wang, “Accurate, scalable in-network identification of p2p traffic using application signatures” World Wide Web 2004, May 17-20, 2004, New York, USA.

[4] Byung-Chul Park, Young J. Won, Myung-Sup Kim, James W. Hong, "Towards Automated Application Signature Generation for Traffic Identification", NOMS 2008, Salvador, Bahia, Brazil, Apri. 7-11, 2008, 160-167.

[5] Patrick Haffner , Subhabrata Sen , Oliver Spatscheck , Dongmei Wang, “ACAS: automated construction of application signatures”, ACM SIGCOMM, August 26-26, 2005, Philadelphia, Pennsylvania, USA.

[6] Xiao, F., Hu, H. “ASG - Automated signature generation for worm-like P2P traffic patterns”. WAIM 2008. July 20-22 2008. pp. 645-660

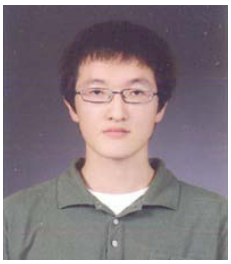
[7] Hun-Jeong Kang, Myung-Sup Kim, and James Won-Ki Hong, "Streaming Media and Multimedia Conferencing Traffic Analysis Using Payload Examination," ETRI Journal, Vol.26, No.3, Jun. 2004, pp.203-217.

[8] F. Risso, M. Baldi, O. Morandi, A. Baldini, and P. Monclus, "Lightweight, Payload-Based Traffic Classification An Experimental Evaluation," IEEE, Beijing, China, May. 19-23, 2008, pp. 5869-5875

[9] Yongmin Choi, "On the Accuracy of Signature-based Traffic Identification Technique in IP Networks" Proc. of the Broadband Convergence Networks, 2007. BcN '07. 2nd IEEE/IFIP International Workshop, Glasgow , Scotland, June. , pp. 1-12

박 준 상
2008 년 고려대학교 컴퓨터정보학과 학사
2008 년~현재 고려대학교 컴퓨터정보학과 석사과정
<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

윤 성 호
2009 년 고려대학교 컴퓨터정보학과 학사
2009 년~현재 고려대학교 컴퓨터정보학과 석사과정
<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석



박 진 완
2009 년 고려대학교 컴퓨터정보학과 학사
2009 년~현재 고려대학교 컴퓨터정보학과 석사과정
<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

이 현 신
2009 년 고려대학교 정보수학과 학사
2009 년~현재 고려대학교 컴퓨터정보학과 석사과정
<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

이 상 우
2010 년 고려대학교 컴퓨터정보학과 학사
<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석,

김 명 섭
1998 년 포항공과대학교 전자계산학과 학사
1998 년~2000 년 포항공과대학교 컴퓨터공학과 석사
2000 년~2004 년 포항공과대학교 컴퓨터공학과 박사
2004 년~2006 년 Post-Doc., Dept. of ECE, Univ. of Toronto, Canada.
2006 년~ 현재 고려대학교 컴퓨터정보학과 조교수
<관심분야> 네트워크 관리 및 보안, 트래픽 모니터링 및 분석, 멀티미디어 네트워크