페이로드 시그니처 기반 트래픽 분석 시스템의 성능 향상

준회원 박 준 상*, 윤 성 호*, 박 진 완*, 이 현 신*, 이 상 우*, 종신회원 김 명 섭*°

Performance Improvement of the Payload Signature based Traffic Classification System

Jun-sang Park*, Sung-ho Yoon*, Jin-wan Park*, Hyun-shin Lee*,
Sang-woo Lee* *Associate Members*, Myung-sup Kim*° *Lifelong Member*

요 약

응용 레벨 트래픽 분석은 네트워크의 효율적인 운영과 안정적인 서비스를 제공하기 위한 필수적인 요소이다. 응용 레벨 트래픽 분석을 위한 다양한 분석 방법이 존재하지만 분류의 정확성, 분석률, 실용성을 고려했을 때 페이로드 시그니처 기반 분석 방법은 가장 높은 성능을 보인다. 하지만 페이로드 시그니처 기반 분석 방법은 고속 링크의 트래픽을 실시간으로 처리하는 과정에서 헤더 정보 및 통계 정보 이용 방법론에 비해 상대적으로 높은 부하를 발생시키며 처리 속도가 느린 단점을 갖는다. 본 논문에서는 페이로드 시그니처 기반 분석 시스템의 처리 속도를 향상시키기 위해 요구되는 디자인 선택 사항을 기술하고, 각 선택 사항에 대해 실험적으로 평가하여 최적화된 분류의 구조를 제시한다. 또한 제안하는 방법을 학내 망에 적용하여 그 타당성을 증명한다.

Key Words : Application-level Traffic Classification; Payload Signature; processing Speed

ABSTRACT

The traffic classification is a preliminary and essential step for stable network service provision and efficient network resource management. While a number of classification methods have been introduced in literature, the payload signature-based classification method shows the highest performance in terms of accuracy, completeness, and practicality. However, the payload signature-based method has a significant drawback in high-speed network environment that the processing speed is much slower than other classification method such as header-based and statistical methods. In this paper, We describes various design options to improve the processing speed of traffic classification in design of a payload signature based classification system and describes our selections on the development of our traffic classification system. Also the feasibility of our selection was proved through experimental evaluation on our campus traffic trace.

Ⅰ. 서 론

네트워크의 고속화와 더불어 다양한 서비스와 응용 프로그램이 개발됨에 따라 기업이나 개인들은 인터넷으로 대표되는 네트워크에 대한 의존이 상당히 커져가고 있다. 이와 같은 현실 속에서 네트워크의 효율적 운용과 관리를 위한 응용 레벨의 트래픽의 모니터링과 분석은 네트워크 사용현황 파악과 확장계획 수립

※ 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 ({junsang_park, sungho_yoon, jinwan_park, hyunshin-lee sangwoo_lee, tmskim}@korea.ac.kr)
(° : 교신저자)
논문번호 : KICS2010-06-269, 접수일자 : 2010년 06월 26일, 최종논문접수일자 : 2010년 8월 12일

등의 다양한 분야에서 필요성이 커져가고 있다. 예를 들어 종량제 과금, CRM, SLA, 보안 분석 등 트래픽 모니터링 및 분석에 대한 필요성은 지금뿐만 아니라 앞으로 더욱더 크게 증가할 것이다. 이를 위해서는 다양한 종류의 응용 레벨 트래픽을 정확하게 분류할 수 있는 방법과 고속 링크에서 발생하는 대용량의 트래픽을 실시간으로 처리하는 방법이 요구된다.

응용 레벨 트래픽 분류 방법에 있어 페이로드 시그니처 기반 분석 방법은 패킷의 헤더 정보나 통계 정보를 이용하는 다른 분석 방법들에 비해 상대적으로 높은 분류 정확성과 분석률을 보인다.[1,3,4,8,9] 하지만 분류 시스템의 처리 속도에 있어 현재의 고속 네트워크 상에서 발생하는 대용량 트래픽을 실시간으로 처리하기에 부적합한 방법이다. 응용의 수와 대용량의 트래픽을 발생시키는 응용의 사용이 증가하고 있는 추세를 량 했을 때 페이로드 기반 분석 방법의 처리 속도 문제는 반드시 해결되어야 하는 과제이다. 따라서 본 논문에서는 페이로드 시그니처 기반 분류 시스템의 처리 속도에 영향을 미치는 요소를 정의하고, 처리 속도 향상을 위한 분류 시스템의 디자인 선택 사항을 실험적으로 평가하여 최적의 분류 시스템 구조를 제시한다.

본 장의 서론에 이어, 2장에서는 페이로드 기반 분석 방법의 문제점에 대해 기술하고, 3장에서는 처리 속도에 영향을 미치는 요소를 평가하기 위한 실험 환경에 대해 기술한다. 4장에서는 실험 결과를 바탕으로 처리 시간 향상을 위한 최적화된 방법을 제안한다. 5장에서는 제안하는 방법을 분류 시스템에 적용하여 그 타당성을 증명한다. 마지막으로 6장에서는 결론 및 향후 연구에 대해 기술한다.

Ⅱ. 관련 연구

응용 프로그램 서비스 제공자는 방화벽을 우회하여 사용자에게 원활한 서비스를 제공하기 위해 복잡한 구조의 응용 레벨 프로토콜 구성하기 때문에 시그니처 또한 다양한 형태로 존재한다. 또한 네트워크 기반의 응용의 증가로 인해 시그니처의 개수가 증가하고 있다. 시그니처의 복잡도가 커지고, 개수가 증가하면서 페이로드 시그니처 기반 분류 시스템의 처리 속도는 트래픽 분류 시스템의 성능을 결정하는 중요한 요소로 작용하게 되었다. 페이로드 시그니처 기반 분류 시스템의 처리 속도를 향상 시키기 위한 연구가 많이 진행되고 있지만 대부분의 연구가 패턴 매칭 알고리즘의 성능을 향상 시키는 부분에 초점을 맞추고 있다.

응용 프로그램 트래픽 분류를 위한 도구로 많이 사용되고 있는 L7-filter는 시그니처를 정규표현식으로 표현하고 패턴 매칭 알고리즘으로 NFA(Nondeterministic Finite Automata)를 적용한다. 하지만 70여 개의 시그니처를 적용하였을 때 3.5Mbps 이하의 처리 속도를 보인다.[6] NFA의 처리 속도를 향상 시키기 위해 DFA(Deterministic Finite Automaton) 기반의 분석이 제안되고 활용되고 있지만 100Mbps 이하의 처리속도를 갖는다.[6,7,11] 분류 시스템의 처리 속도 향상을 위해 패턴 매칭 알고리즘의 성능 향상을 위한 방법을 제안하지만 매칭 알고리즘의 성능은 입력 데이터의 구성에 의존적이며, 제한적인 성능 향상을 나타낸다.

우리는 선행 연구[1]를 통하여 126개의 응용프로그램에 대해 845개의 페이로드 시그니처를 추출하였다. 추출한 시그니처를 학내 망의 전체 인터넷 트래픽에 대해 실시간으로 적용한 결과 flow/byte/packet 단위로 99%이상의 정확도와 85% 이상의 분석률을 보였다. 하지만 표 1과 같은 시스템 환경에서 최대 처리 가능한 처리 속도는 160Mbps로 나타났다. 그림 1은 선행 연구의 분류 시스템의 최대 처리 속도와 학내 망에서 발생하는 트래픽의 bandwidth를 나타내고 있다. 기존의 분류 시스템은 학내 망의 트래픽을 실시간으로 처리하지 못하는 문제점을 보였다.

패킷 헤더 정보에 기반한 분류 방법과 비교를 통해서도 기존의 페이로드 시그니처 기반 분석 방법의 처리 시간 문제점을 확인할 수 있었다.

표 2는 1일 동안 학내 망과 인터넷 사이에서 발생한 트래픽의 양을 나타내며, 표3은 표2의 트래픽을 1분 단위로 패킷의 헤더 정보만을 매칭하는 고정

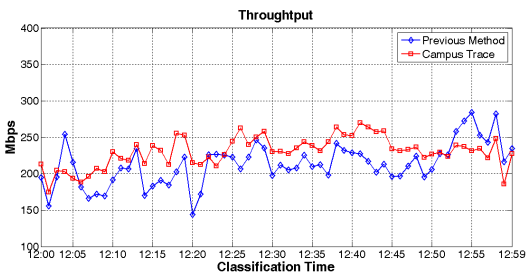

그림 1. 기존 분류 시스템의 처리량

표 1. 분류 시스템 구성

CPU	Intel® Core2 Duo E7200 2.53GHz
Memory	3GByte
O/S	Linux Cent OS

IP-port 기반 분석 시스템[5]과 페이로드 기반 분석 시스템으로 분류하는 수행 시간을 비교한 결과이다. 페이로드 기반 분석 방법과 고정 IP_port 기반 분석 방법은 시그니처의 매칭 과정을 제외하고 동일한 프로세싱을 통해 분류된다.

표 3 에서 확인 할 수 있듯이 페이로드 시그니처 기반 분석 방법은 헤더 정보에 기반한 방법에 비해 많은 처리 시간이 요구된다. 이는 두 가지 방법이 동일한 처리 과정으로 분류되는 것을 고려했을 때 페이로드와 시그니처의 매칭 과정의 복잡성 때문으로 판단된다. 따라서 본 논문에서는 시그니처 매칭 과정의 복잡성에 영향을 미치는 요소를 정의하고 최적화된 분류 시스템 구조를 제시하고자 한다.

표 2. 트래픽 양

단위	Flows(K)	Packets(M)	Bytes(GB)
Volume	52,282	6,051	58,334

표 3. 트래픽 분석 시간 비교

(단위: msec)

Method	Min.	Max.	Avg.
Fixed IP-port	24	210	81
Payload	1,252	45,765	17,235

Ⅲ. 분류 시스템 및 실험 환경

본 장에서는 분류 시스템의 성능을 객관적으로 평가하기 위한 분류 및 검증 시스템의 구성과 실험 환경에 대해 기술한다.

3.1 분석 및 검증 시스템의 구성

그림 2는 분류 시스템의 성능 평가를 위해 구성한 트래픽 수집, 분류, 검증 시스템인 KU-MON 시스템의 구성을 나타내고 있다. TCS를 통해 학내 망과 인터넷을 연결하는 링크의 모든 패킷을 수집하여 bidirectional-flow로 구성하고 수집된 플로우는 TAS로 전송되어 페이로드 시그니처를 기반으로 응용 단위로 분류된다.[1]

결과의 정확성을 검증하기 위해 TMA[2] 기반으로 Ground Truth 트래픽을 수집한다. TMA는 학내망의 단말 호스트에 설치되며 소켓 정보를 기반으로 하여 Process name, IP, port, protocol, path 등의 정보를 생성한다. TMA가 설치된 호스 트에서 열려진 소켓을 주기적으로 검사하여 TMS로 TMA정보를 전송하고

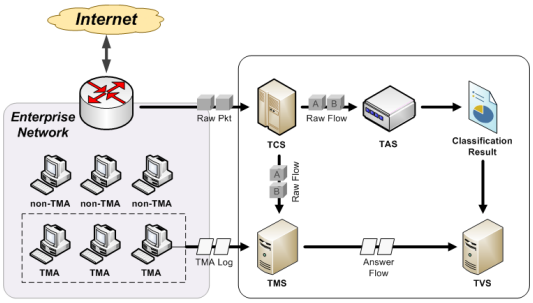

그림 2. 분류와 검증 네트워크의 구성

TMS는 각 호스트로부터 전달받은TMA 정보를 통합하여 분류 시스템의 분류 결과의 Ground Truth를 제공한다. TVS에서는 TAS의 결과와 Ground Truth를 비교하여 성능 평가를 실시한다.

3.2 트래픽 트래이스

표 4는 실험에 사용된 트래픽 트래이스와 실험 기간을 보여주고 있다. 트래픽은 학내 망과 인터넷의 연결 지점에서 수집한 데이터로 3000여대의 호스트에서 사용하는 다양한 종류의 응용 프로그램의 트래픽으로 구성된다.

표 4. 트래픽 트래이스

	Flow	Packet	Byte
	20,489,392	773,384(K)	701,751(MB)
기간	2009.10.17 12:00 ~ 2009.10.17 12:59		

3.3 분류 시스템의 구성

표 5는 평가의 기준이 되는 분류 시스템의 구성을 나타내고 있다. 응용 프로그램을 기준으로 분류하며, 응용 프로그램은 해당 응용 프로그램이 동작시키는 프로세스의 집합으로 구성된다. 패킷은 플로우 단위로 수집되며 분류된다. 플로우는 분류의 정확성을 향상시키고, reverse-flow에 대한 처리가 용이하도록 bidirectional 형태로 구성된다. 매칭 알고리즘은 분류의 정확성을 확인하는 지표로 사용하기 위해 전사적 방법으로 매칭하는 Brute-force 알고리즘을 선택하였다. 시그니처는 패킷 단위로 매칭되며 분석된다. 분류 시스템은 1분 주기로 수집된 플로우를 대상으로 동작한다.

표 5의 분류 시스템의 구성에 기초하여 분류 시스템의 처리 속도에 영향을 미치는 요소를 조사하고, 처리 속도 향상을 위한 최적화된 분류 시스템의 구조를

표 5. 분류 시스템의 구성

분류 범위	Process	Application	Signature
	260	126	845
분류 기준	Application(Set of Process)		
분류 단위	Bidirectional Flow		
분류 알고리즘	Brute-force		
매칭 단위	Packet		
분류 시간 단위	1min		

설계한다.

실험에 사용된 분류 시스템은 표 1과 같은 범용 컴퓨터의 환경으로 구성되어 있고, 본 논문에서는 이와 같은 환경에서 최대의 성능을 보이는 분류 시스템을 구성하는 것을 목표로 하고 있다.

Ⅳ. 분류 시스템 디자인 선택 사항

본 장에서는 분류 시스템의 처리 속도에 영향을 미치는 요소를 정의하고, 처리 속도를 향상 시킬 수 있는 최적의 선택 사항을 제시한다.

본 논문에서는 분류 시스템의 처리 속도에 영향을 미치는 요소를 입력 데이터의 탐색공간 크기와 매칭 알고리즘의 성능을 구분하여 기술한다.

4.1 입력 데이터의 탐색 공간

분류 시스템의 입력 데이터 탐색 공간을 최소화하여 처리 속도를 향상 시킬 수 있는 방법에 대해 기술한다. 분류 시스템은 분석 대상 플로우와 시그니처를 입력 데이터로 제공받는다. 분석 시스템은 하나의 플로우를 분류하기 위해 모든 시그니처 매칭 과정을 통해 분류하기 때문에 플로우 내의 조사되는 패킷과 시그니처에 대한 탐색 공간은 처리 속도에 절대적인 영향을 미친다. 따라서 조사하는 패킷과 시그니처의 탐색 공간을 줄임으로서 처리 속도를 향상 시킬 수 있다. 하지만 이는 분류의 정확성과 분석률에 직접적인 영향을 미치는 요소이기 때문에 정확성과 분석률에 대한 정확한 검증 결과가 수반되어야 한다.

4.1.1 시그니처의 중복성 제거

동일한 응용 프로그램 내의 두 개 이상의 시그니처가 하나의 플로우를 중복 분류하는 과정을 피해야 한다. 시그니처의 개수가 증가하면서 동일한 응용 프로그램 내의 시그니처가 포함 관계를 갖거나 중복된 형태로 등록되는 경우가 발생한다. 본 논문에서는 시그니처 등록 전에 중복성을 확인하여 시그니처의 탐색 공간을 줄이는 모듈을 구성하였다.

4.1.2 시그니처의 계층 구조

응용 프로그램 시그니처는 그림 3과 같이 응용 레벨 프로토콜 시그니처와 응용 프로그램 시그니처의 2단계 계층 구조로 표현 할 수 있다. 시그니처의 계층 구조는 시그니처의 포함 관계를 기반으로 정의된다. 시그니처 S_x에 의해 분류되는 트래픽이 시그니처 S_y에 의해 모두 분류되면 S_y는 S_x를 포함한다. 이때 S_y는 S_x의 응용 프로토콜 레벨 시그니처가 되며, S_x는 S_y의 응용 프로그램 레벨 시그니처가 된다.

표 6은 계층 구조를 갖는 시그니처의 개수를 나타낸다. 상위 계층 시그니처가 62개이며, 상위 계층에 포함되는 시그니처가 129이다. 나머지 시그니처는 포함 관계를 갖는 않는 시그니처로 모든 플로우에 항상 비교되는 시그니처이다.

그림 4는 계층 구조 기반 분석 방법에서 1개의 플로우를 분류하기 위해 사용되는 평균 시그니처의 개수를 나타낸다. 평면적 구조의 기존의 방법은 1개의 플로우를 분류하기 위해 845개의 시그니처를 모두 분류해야 하지만 계층 구조 기반 분석 방법은 평균적으로 100여개 이상의 시그니처의 탐색 공간이 감소되는 것을 확인 할 수 있다.

그림 5는 시그니처를 평면 구조와 계층 구조로 표현하고 분류한 수행 시간을 나타낸다. 두 가지 방법 모두 트래픽의 양과 응용 프로그램의 수가 증가 할수록 처리시간이 길어지는 것을 알 수 있다. 하지만 평면적 분석 방법에 비해 계층적 구조의 분석 방법이 평균적으로 1.5배 정도 빠른 것을 알 수 있다. 이는 분류 시스템에서 처리해야 하는 시그니처의 탐색 공간이 감소했기 때문이다. 평면 구조와 계층 구조 기반 분석

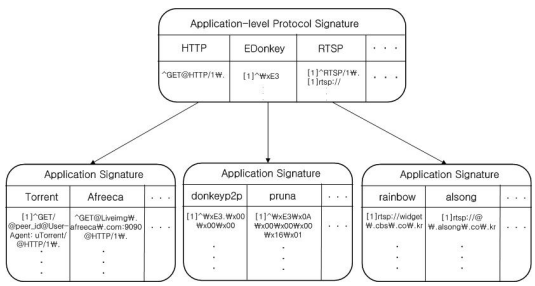

그림 3. 시그니처 계층 구조

표 6. 시그니처 계층 구조 통계

Total Sig	App Prot. Sig	App Sig	Others
845	62	129	654

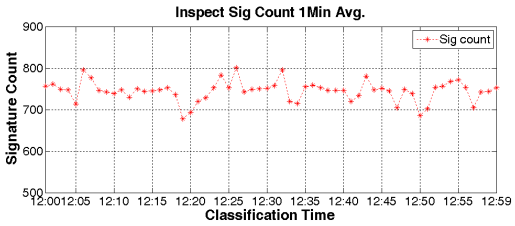

그림 4. 검사하기 위한 시그니처 개수

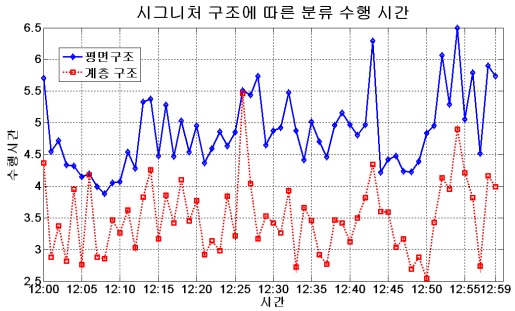

그림 5. 시그니처 구조에 따른 수행 시간 비교

방법의 분류 시간이 유사하게 나타나는 부분은 그림 4에서 알 수 있듯이 계층 구조 기반 분석 방법에서 비교하는 시그니처 개수가 평면 구조와 유사하게 나타나는 경우이다. 이는 응용 프로토콜 레벨의 시그니처에 의해서 분류되는 빈도가 적기 때문이다.

4.1.3 조사하는 패킷 개수

플로우 내의 조사하는 패킷의 개수를 제한함으로써 시그니처 탐색 시간을 감소 시킬 수 있다. 표 7은 조사되는 플로우의 초기 패킷 개수를 증가 시키면서 분석률과 정확도를 측정한 결과이다. 이 때 첫 번째 패킷은 TCP연결 설정 이후 페이로드가 존재하는 최초의 패킷으로 정의한다. 패킷 개수의 제한에 따른 분석 결과 분류의 정확성과 분석률은 3번째 패킷 이후에는 동일한 것을 알 수 있다. 이는 페이로드 시그니처가 서버-클라이트 사이에서 컨트롤 패킷을 전송하는 과정에서 추출되기 때문이다. 대부분의 플로우는 콘텐츠를 전송하기 전에 응용 레벨 프로토콜에 의해서 약속된 컨트롤 패킷을 전송하고 컨텐츠를 전송한다. 따라서 3번째 패킷 이내에서 분류될 수 있다.[10]

분류된 트래픽에 대한 분류 정확도는 플로우, 바이트, 패킷 모두99% 이상을 나타내지만 패킷과 바이트의 분석률이 낮은 원인은 실험을 수행한 시점에 분류 범위에 포함되지 Web-Disk 형태의 파일 공유 프로그램의 트래픽이 발생하였기 때문이다. Web-Disk 형태의 응용 프로그램은 적은 수의 플로우로 대용량의 트래픽을 발생시키기 때문에 플로우단위의 분석률보다는 패킷과 바이트에 크게 영향을 미친다.

그림 6은 조사하는 플로우의 초기 패킷 개수를 1-5로 증가 시켰을 때의 수행 시간을 나타내고 있다. 그림 6에서 알 수 있듯이 조사하는 패킷의 개수의 증가는 처리 속도에 직접적인 영향을 미치기 때문에 분석률와 정확도에 영향을 주지 않는 3개의 패킷으로 제한한다.

표 7. 패킷 개수에 따른 분석률과 분류 정확도

(단위 : %)

		Pkt1	Pkt2	Pkt3	Pkt4	Pkt5
분석률	Flow	91.5	91.6	91.6	91.6	91.6
	Pkt	40.4	40.6	41.2	41.2	41.2
	Byte	35.2	35.7	36.2	36.2	36.2
정확도	Flow	96.4	96.9	96.9	96.9	96.9
	Pkt	98.5	98.7	98.7	98.7	98.7
	Byte	99.7	99.8	99.8	99.8	99.8

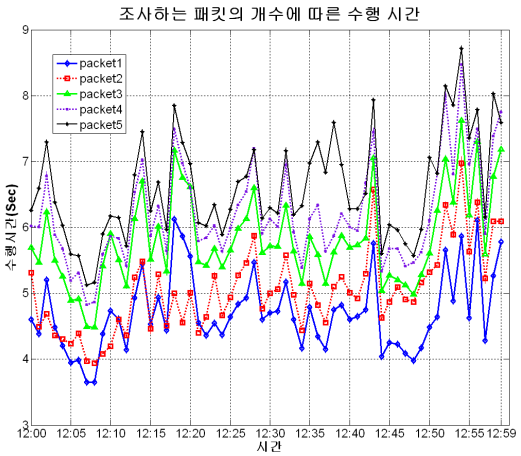

그림 6. 패킷 개수에 따른 수행 시간 비교

4.1.4 시그니처의 표현

본 논문에서는 응용프로그램의 페이로드 시그니처를 정규표현식으로 기술하고, 처리 속도 향상을 위해 패킷 순서 정보를 표현하는 방법을 추가하였다. 표 8은 국내에서 주로 사용되는 P2P 응용 프로그램으로 알려진 소리바다에 대한 페이로드 시그니처를 나타낸

표 8. 시그니처 기술

시그니처
[1]^GET.*Soribada.*HTTP/1\.1.*www\.soribada\.com
[1]soribada\.com
[1]^GETMP3\x0D\x0AFilename

다. 시그니처는 정규 표현식으로 기술되며, 각 시그니처에 '['과 ']' 사이에 패킷의 순서 정보를 명시함으로써 분류 시스템은 플로우 내의 해당 패킷만을 조사할 수 있다. 따라서 분류 시스템에서 패킷에 대한 탐색 공간을 줄여 처리 속도를 향상 시킬 수 있다.

4.2 패턴 매칭 알고리즘의 처리 속도 향상

패턴 매칭 알고리즘은 분류 시스템의 처리 속도에 가장 큰 영향을 미치는 요소이다. 본 논문에서는 패턴 매칭 분야에서 성능이 검증된 4개의 알고리즘에 대해 실험적으로 평가하고 응용 레벨 트래픽 분석에 적합한 알고리즘을 제시한다.

4.2.1 패턴 매칭 알고리즘

표 9는 실험에 사용한 패턴 매칭 알고리즘과 각 알고리즘의 시간 복잡도를 보여주고 있다. 표 9의 알로리즘 외에도 다양한 스트링 매칭 알고리즘이 존재하지만 매칭 알고리즘의 시간 복잡도가 낮은 알고리즘을 선택하였다.[12]

Brute-force 알고리즘은 시그니처를 패킷 페이로드에 1바이트씩 오른쪽에서 왼쪽으로 이동하면서 모든 경우에 대해 매칭하는 방법이다. 이러한 방법은 전처리 과정에 불필요하고 매칭 오류가 발생하지 않기 때문에 정확도의 기준으로 사용될 수 있다. 하지만 처리 속도 측면에서는 가장 성능이 낮은 방법이다.

DFA(Deterministic Finite Automata)는 각각의 시그니처마다 1개의 유한 오토마타로 구성하고 패킷의 페이로드와 매칭하는 알고리즘이다. 이는 시그니처를 오토마타로 구성하기 위한 preprocessing 과정이 요구된다.

Robin-karp 알고리즘은 시그니처에 대한 해쉬 값을 구하고, 페이로드에서 1바이트씩 왼쪽에서 오른쪽으로 이동하면서 해쉬 값을 비교하여 값이 동일한 경우 다시 매칭하는 구조를 갖는다. 따라서 해쉬 함수에 의해 처리 속도가 영향을 받는다.

Boyer-Moore 알고리즘은 패턴의 일부가 매칭되면 일치되는 정보에 대한 위치 정보를 기억하여 이동하는 윈도우의 크기를 최대로 함으로써 매칭 속도를 향상 시키고자 하는 알고리즘이다. 이때 윈도우는 왼쪽에서 오른쪽으로 이동하며 패턴에 대한 검사는 오른쪽에서 왼쪽 방향으로 진행된다. 패턴의 일부가 페이로드 내에서 나타나는 빈도가 높은 경우 효율적인 방법이지만 반대의 경우 Brute-force 알고리즘과 동일한 성능을 나타내는 단점이 있다.

그림 7은 표 9의 4가지 알고리즘에 대해 패킷 개수를 3개로 제한하여 매칭 알고리즘의 처리 속도를 측정한 결과이다. 각각의 알고리즘의 분류 정확성은 동일한 결과로 나타났다.

분류 시스템은 하나의 입력 플로우에 대해 전체 시그니처를 모두 비교하는 과정이 요구된다. 하지만 전체 시그니처 중 1-2개의 시그니처만 패턴 일치가 발생하고 대부분의 시그니처는 패킷의 페이로드의 끝까지 조사해야 한다. 따라서 전체 페이로드에 대한 비교시간이 빠른 알고리즘이 요구된다. Robin- Karp 알고리즘은 해시 값으로 전체 페이로드를 빠르게 비교할 수 있기 때문에 가장 높은 성능을 보인다. Boyer-Moore 알고리즘은 시그니처의 일부 문자나 문자열이 페이로드를 구성하는 문자열에 일치하는 빈도가 잦은 경우 빠른 처리 속도를 나타낸다. 하지만 페이로드의 특성상 반복되는 문자의 발생 빈도가 적기 때문에 부적합한 알고리즘이다. DFA는 각 시그니처를 오토마타로 구성하는 Preprocessing 시간의 지연으로 Robin-Karp 알고리즘에 비해 느린 처리 속도를 보였다.

표 9. 매칭 알고리즘과 시간 복잡도

Algorithm	Preprocessing	Matching
Brute-Force	No preprocessing	Θ(nm)
DFA	Θ(m |Σ|)	Θ(n)
Robin-Karp	Θ(m)	Θ(n+m),
Boyer-Moore	Θ(m + |Σ|)	Θ(n)

n : 패킷 페이로드 길이
m : 시그니처 길이
Σ : 나타날 수 있는 문자의 경우의 수

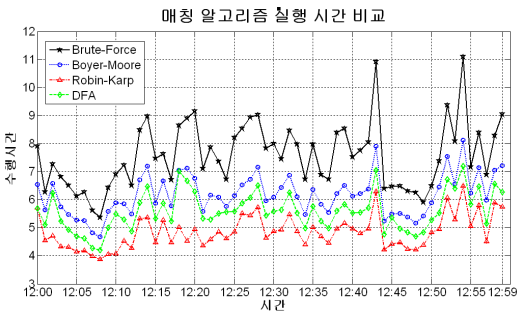

그림 7. 매칭 알고리즘의 실행 시간 비교

V. 성능 평가

본 장에서는 4장에서 기술한 분류 시스템의 처리 속도에 영향을 미치는 요소를 적용하여 성능을 평가

한다. 표 10은 각 요소에 대해 제안하는 방법을 정리한 결과이다.

표 10의 내용을 분류 시스템에 적용하고 최대 처리 가능한 bps를 통해 성능을 평가하였다. 측정 결과 기존의 방법에 비해 평균적으로 240Mbps의 처리 속도가 향상되었다.

성능 평가 결과 패턴 매칭 알고리즘에 의한 처리 속도 향상은 30-50Mbps로 크게 영향을 미치지 않았다. 처리 속도에 대한 성능 향상은 입력 데이터의 탐색 공간을 줄이는 부분에서 가장 크게 나타났다. 기존에 10개의 패킷을 모두 탐색하는 방법에서 3개의 패킷에 대한 탐색만으로 정확한 분류가 가능하고, 계층 구조 기반 분석을 통해 100개 이상의 시그니처의 탐색 공간을 감소 시킬 수 있기 때문이다. 트래픽의 양에 따른 처리 시간의 변화를 확인한 결과 처리 시간은 플로우의 개수 보다는 패킷과 바이트의 양에 따라 다른 처리 시간을 보였다. 이는 패킷이 많고 바이트가 큰 경우 플로우 내에 조사하는 패킷의 개수가 많아지고, 패킷 내에서 조사하는 페이로드의 탐색 공간이 늘어나기 때문이다.

표 10. 제안하는 방법의 구성

입력 데이터 최적화	시그니처 매칭 구조	계층 구조기반
	패킷 개수	3
	패킷 순서정보	시그니처에 포함
패턴 매칭 알고리즘		Robin - karp

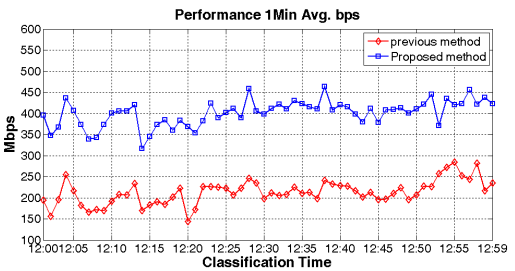

그림 8. 기존의 방법과 제안하는 방법의 처리 속도 비교

VI. 결론 및 향후 과제

본 논문에서는 페이로드 기반 응용 레벨 트래픽 분류 시스템의 처리 속도에 영향을 미치는 요소들을 탐색 공간과 매칭 알고리즘의 처리 속도로 구분하여 각 요소를 실험적으로 평가하고 효율적인 분류 시스템을 구성하기 위한 방법을 제안하였다.

본 논문에서는 분류 시스템의 처리 속도를 향상시키기 위해 입력 데이터의 탐색 공간과 패턴 매칭 알고리즘을 단일 프로세서 환경에서 적용하였다. 향후 연구로 제안하는 방법을 다중 프로세스 환경에 적합한 구조로 변경하는 연구와 기존의 패턴 매칭 알고리즘을 보완하여 페이로드 시그니처에 최적화된 매칭 알고리즘을 설계할 계획이다.

참 고 문 헌

〔1〕 박준상, 박진완, 윤성호, 오영석, 김명섭, “응용 레벨 트래픽 분류를 위한 시그니처 생성 시스템 및 검증 네트워크의 개발”, 제31회 정보처리학회 춘계학술발표대회, 부산, 한화리조트, Apr. 23-24, 2009, 제16권 제1호, pp.1288-1291.

〔2〕 윤성호, 노현구, 김명섭, “TMA(Traffic Measurement Agent)를 이용한 인터넷 응용 트래픽 분류”, 통신학회 하계종합학술발표회, 라마다플라자호텔, Jul. 2-4, 2008, pp.618.

〔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〕 F. Risso, M. Baldi, O. Morandi, A. Baldini, and P. Monclus, “Lightweight, Payload-Based Traffic Classification An Experimental Evaluation,” IEEE International Conference on Communications, Beijing, China, May. 19-23, 2008, pp. 5869-5875.

〔5〕 Sung-Ho Yoon, Jin-Wan Park, Young-Seok Oh, Jun-Sang Park, and Myung-Sup Kim, “Internet Application Traffic Classification Using Fixed IP-port,” APNOMS 2009, LNCS, Jeju, Korea, Sep. 23-25, 2009, pp.21-30.

〔6〕 Fnag Yu, Zhifeng Chen, Yanlei Dino, T. V. Lakshman, Randy H. Katz, “Fast and memory Efficient Regular Expression Matching for Deep Packet Inspection” ANCS 2006, December , 2006, San jose, California USA.

〔7〕 Christopher L. Hayes , Yan Luo, “DPICO: a high speed deep packet inspection engine using compact finite automata”, ACM/IEEE Symposium on Architecture for networking and communications systems, December 03-04,

2007, Orlando, Florida, USA
〔8〕 Liu, Hui Feng, Wenfeng Huang, Yongfeng Li, Xing "Accurate Traffic Classification", Networking, Architecture, and Storage, NAS 2007. International Conference
〔9〕 Byung-Chul Park, Young Won, Mi-Jung Choi, Myung-Sup Kim, and James W. Hong, "Empirical Analysis of Application-Level Traffic Classification Using Supervised Machine Learning," Proc. of the Asia-Pacific Network Operations and Management Symposium (APNOMS) 2008, LNCS5297, Beijing, China, Oct. 22-24, 2008, pp.474-477.
〔10〕 Yuhai Liu, Hongbo Liu, Hongyu Zhang, Xin Luan, "The Internet Traffic Classification an Online SVM Approach", ICOIN 2008. International Conference, Busan, Korea, Jan. 23-25, 2008, pp.1-5.
〔11〕 G. Vasiliadis, M. Polychronakis, S. Antonatos, E. P. Markatos, and S. Ioannidis, "Regular expression matching on graphics hardware for intrusion detection," in RAID, 2009, pp.265-283.
〔12〕 Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, Clifford Stein. Introduction to Algorithms, Second Edition. MIT Press and McGraw-Hill, 2001. ISBN 0-262-03293-7. Chapter 32: String Matching, pp.906-932.

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 현 신 (Hyun-shin Lee) 준회원

2009년 고려대학교 정보수학과 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) 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