http://dx.doi.org/10.3745/KIPSTC.2011.18C.4.243

통계 시그니쳐 기반 트래픽 분석 시스템의 성능 향상

박 진 완[†] · 김 명 섭[††]

요 약

네트워크의 고속화와 다양한 서비스의 등장으로 오늘날의 네트워크 트래픽은 복잡 다양해지고 있다. 효율적인 네트워크 관리를 위해서는 네트워크에서 발생하는 트래픽에 대한 다양한 분석이 필요하다. QoS, SLA와 같은 정책을 적용하기 위해서는 트래픽 분석 중에서도 트래픽 분류의 중요성이 크다. 현재까지 트래픽 분류에 관한 연구가 활발히 진행되어 왔는데 최근에는 플로우의 통계 정보를 이용한 트래픽 분류 방법론이 많이 연구되고 있다. 본 논문에서는 기존 연구에서 제안한 페이로드 크기 분포를 이용한 트래픽 분류 방법의 문제점인 낮은 분석률 및 정확도를 향상시키는 방법을 제안한다. 본 논문에서 제안하는 방법은 PSD 충돌로 인해 분류하지 못하는 트래픽을 IP와 port정보를 이용하여 추가적으로 분류하여 분석률을 향상시키고 기존 분류 방법에서 트래픽 분류를 위해 사용되던 플로우와 시그니쳐 사이의 거리 측정 방법을 벡터 거리 측정에서 패킷 별 거리 측정으로의 변경으로 통해 분류 방법의 정확도를 향상시킨다. 제안한 방법은 학내 망에서의 실험을 통해 기존 알고리즘에 비해 향상된 알고리즘의 성능을 검증한다.

키워드 : 트래픽 분류, 트래픽 분석, 통계 시그니쳐, 응용 트래픽

Performance Improvement of the Statistic Signature based Traffic Identification System

Jin-Wan Park[†] · Myung-Sup Kim[††]

ABSTRACT

Nowadays, the traffic type and behavior are extremely diverse due to the appearance of various services on Internet, which makes the need of traffic identification important for efficient operation and management of network. In recent years traffic identification methodology using statistical features of flow has been broadly studied. We also proposed a traffic identification methodology using payload size distribution in our previous work, which has a problem of low completeness. In this paper, we improved the completeness by solving the PSD conflict using IP and port. And we improved the accuracy by changing the distance measurement between flow and statistic signature from vector distance to per-packet distance. The feasibility of our methodology was proved via experimental evaluation on our campus network.

Keywords : Traffic Identification, Traffic Analysis, Statistic Signature, Application Traffic

1. 서 론

네트워크의 고속화와 다양한 서비스의 등장으로 오늘날의 네트워크 트래픽은 복잡 다양해지고 있다. 이러한 상황 속에서 효율적인 네트워크 관리를 위해 트래픽 분석의 중요성을 점점 증가될 전망이다[1, 2]. QoS(Quality of Service), SLA(Service Level Agreement), CRM(Customer Relationship Management)과 같은 정책을 적용하기 위해서는 트래픽 분석 중에서도 트래픽 분류의 중요성이 크다. 따라서, 예전부터 트래픽 분류에 관한 연구는 활발히 진행되어 왔다. 많은 분류 방법론이 개발되었지만, 최근에는 플로우의 통계 정보를 이용한 트래픽 분류 방법론[3, 4, 5]이 많이 연구되고 있는 실정이다.

플로우의 통계 정보를 이용한 분류 방법은 패킷 크기, 패킷 간의 시간 간격, 윈도우 크기 등 패킷들로부터 얻어지는 다양한 통계적 특징을 이용하여 머신 러닝의 특정 알고리즘들을 사용하여 트래픽을 분류하는 방법이 주로 제안되어 왔다[6]. 또한, 특정 통계적 정보를 이용하여 자체적인 알고리즘을 개발한 연구들도 진행되었는데, 그 중 패킷 또는 페이로드 크기 분포를 이용한 분류 방법들[4, 7, 8, 9, 10]이 많이 제안되고 높은 정확도를 나타내었다.

※ 이 논문은 2009년 정부(교육과학기술부)의 재원으로 한국연구재단(2009-0090455)의 지원을 받아 수행된 연구임.
† 정 회 원 : 고려대학교 컴퓨터정보학과 석사
†† 종신회원 : 고려대학교 컴퓨터정보학과 부교수(교신저자)
논문접수 : 2011년 1월 24일
수 정 일 : 1차 2011년 4월 25일
심사완료 : 2011년 5월 17일

본 논문에서는 약 6개월 동안 수집한 정답지 트래픽을 분석을 통하여 기존 페이로드 크기 분포를 이용한 트래픽 분류 방법의 분류 한계점을 파악하고, 그 한계점을 극복하기 위해 추가적인 트래픽 특징으로써 IP와 port를 이용하는 방법을 제시한다.

제안한 방법을 통해 기존 연구인 통계 시그니쳐 기반 트래픽 분류 방법론의 문제점인 낮은 분석률을 극복하고, 더불어 분류에 사용되는 거리 측정 방법의 변경을 통해 분류 정확도를 향상시킨다. 개발한 알고리즘은 학내 망에 분석 시스템으로 구현하고 검증을 통해 실효성을 증명한다.

본 논문은 다음과 같이 구성된다. 서론에 이어 2장에서는 관련 연구로써 기존 통계 시그니쳐 기반 분류 방법에 대해 간략히 설명한다. 3장에서는 정답지 트래픽 분석을 통해 페이로드 크기 분포의 분류 한계점 및 극복 방안에 대해 설명한다. 4장에서는 통계 시그니쳐를 이용한 트래픽 분류 방법에 대해 설명한다. 5장에서는 제안한 방법론의 우수성을 검증하기 위해 학내 망 트래픽에 대한 실험과 분석 결과를 기술한다. 마지막으로 6장에서는 결론 및 향후 연구에 대해 기술한다.

2. 관련 연구

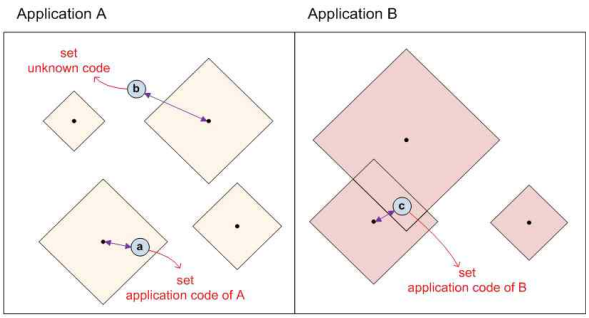
본 장에서는 이전 논문[11]에서 제안한 통계 시그니쳐 기반 방법론에 대해 간략히 설명한다.

통계 시그니쳐란 패킷의 헤더 정보(패킷 크기, 윈도우 크기 등)나 캡쳐 정보(패킷 캡쳐 시간 등)를 기반으로 하여 다른 응용 프로그램과 구별할 수 있는 응용 프로그램의 고유한 통계적 특징을 의미한다. [11]에서는 여러 가지 통계적 특징 중에서 페이로드 크기 분포를 시그니쳐로 생성하고 이를 통해 트래픽을 분류하는 방법을 제안하였다.

[11]에서 사용하는 페이로드 크기 분포(Payload Size Distribution(PSD))란 양방향 플로우에서 첫 N개 데이터 패킷의 페이로드 크기와 방향을 의미한다. 이 때, 페이로드를 포함하지 않는 TCP 컨트롤 패킷(SYN, RST, FIN, …) 등은 N개에 포함되지 않는다. 페이로드 크기는 데이터 패킷(페이로드가 존재하는 패킷)의 페이로드 크기를 의미한다. 방향은 양수와 음수로 표현되며, TCP의 경우 양수는 클라이언트에서 서버로 향하는 패킷, 음수는 서버에서 클라이언트로 향하는 패킷을 의미한다. UDP는 서버/클라이언트의 구분이 명확하지 않기 때문에, 양수/음수의 의미는 단지 방향이 서로 반대라는 것만 표현할 수 있다. 따라서, UDP의 경우에는 발생되는 첫 패킷을 양수로 표현하고 뒤에 이어지는 패킷은 첫 패킷을 기준으로 방향이 같으면 양수, 다르면 음수로 표현한다.

양방향 플로우는 PSD를 나타내는 최대 N차원의 벡터로 표현되며, 이를 PSD 벡터 또는 PSD 패턴이라 지칭한다. 통계 시그니쳐는 응용 프로그램 이름, 전송 계층 프로토콜, PSD 벡터, 거리 임계값으로 표현된다.

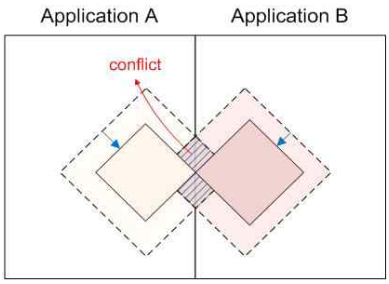
통계 시그니쳐를 이용한 트래픽 분류 방법은 다음과 같다.



(그림 1) 통계 시그니쳐 기반 트래픽 분류의 예시

분류하고자 하는 플로우를 PSD 벡터로 표현하고, 이를 시그니쳐의 PSD 벡터와의 유사성을 통해 분류한다. 유사성은 벡터 간의 거리로 측정되며, 두 벡터 간의 거리가 시그니쳐의 거리 임계값보다 작거나 같으면 해당 시그니쳐의 응용 프로그램 이름으로 분류한다.

(그림 1)은 기존 연구[11]에서 사용하는 통계 시그니쳐 기반 트래픽 분류 방법의 예시를 보여준다. 마름모는 통계 시그니쳐를 나타내며, 중심값인 PSD 벡터와 마름모의 크기인 거리 임계값에 의해 표현된다. 그림에서는 세 개의 플로우 a, b, c에 대한 분류 결과를 보여준다. 플로우 a는 응용 A의 시그니쳐 중 하나의 시그니쳐에 의해 분류되는 경우로써, 시그니쳐의 PSD 벡터와의 거리가 거리 임계값 이하이므로 응용 A로 분류한다. 플로우 b는 어떠한 시그니쳐에도 속하지 않는 경우로써, 미확인(unknown) 트래픽으로 분류한다. 플로우 c는 같은 응용 내에서 두 개 이상의 시그니쳐에 속하는 경우이며, 예시에서는 응용 B로 분류된다.



(그림 2) 통계 시그니쳐 생성 단계에서의 충돌 처리

(그림 2)는 통계 시그니쳐 생성 단계 중 한 단계인 충돌 처리에 대해 보여준다. 시그니쳐 생성에서 서로 다른 응용 간의 시그니쳐가 겹치는 경우를 충돌이라고 정의한다. 분류 시 충돌 영역에 속하는 플로우는 2개 이상의 응용에 의해 분류 가능하므로, 이 중 하나의 응용으로 분류할 경우 잘못 분류할 가능성이 있다. 따라서, 기존 연구[11]의 시그니쳐 생성 단계에서는 이러한 충돌을 제거하기 위해 충돌을 발생시키는 시그니쳐의 거리 임계값을 줄인다. 즉, 잘못 분류할 가능성이 있는 트래픽을 분류하지 않음으로써 분류 정확도를 높인다. 그러기 때문에 기존 연구의 문제점으로 낮은 분석률이 제기되었다.

본 논문에서는 기존 연구의 낮은 분석률을 해결하기 위해 충돌 영역에 존재하는 플로우들을 추가적으로 IP와 port정보를 사용하여 분류한다. 기존 연구의 분류 정확도는 97% 이상의 높은 정확도를 보였지만, 좀 더 높은 정확도를 위해 본 논문에서는 플로우와 시그니쳐의 거리 측정 방법을 벡터 거리 측정에서 패킷 별 거리 측정으로 변경한다.

3. PSD의 분류 한계점과 극복 방안

본 장에서는 PSD의 분류 한계점인 PSD 충돌에 대해 살펴보고 이에 대한 극복 방안을 제시한다. 이는 정답지(ground-truth) 트래픽에 대한 분석을 통해 이루어진다.

3.1 정답지 트래픽 트레이스

트래픽 분류에 사용되는 시그니쳐를 생성하거나 분류 규칙을 생성하기 위해서는 정답지(ground-truth) 트래픽이 반드시 필요하다. 이러한 정답지는 매우 정확해야만 시그니쳐의 신뢰성을 보장해준다. 본 연구에서는 TMA-에이전트 기반의 정답지 생성 방법[12]을 이용하여 정확한 정답지를 학내 망에서 지속적으로 얻어낸다. <표 1>은 PSD 분석에 사용된 정답지 트래픽에 관한 정보이다.

〈표 1〉 정답지 정보

	Process	Flow ($\times 10^3$)	Packet ($\times 10^6$)	Byte ($\times 10^9$)
TCP	431	2,828	697	534
UDP	62	552	18	16
Total	446	3,380	715	550

약 6개월(2009년 7월 6일 4시 1분 ~ 2010년 1월 12일 14시 1분) 동안 정답지 트래픽을 수집하였다. 학내 망의 261개의 호스트에서 정답지가 추출되었으며, 총 446개의 프로세스에서 총 약 338 만개의 플로우가 얻어졌다. 실험한 시점에서 서비스가 없어진 응용이나 잘못 추출된 응용은 정답지에서 제외하여 실험에 사용하였다.

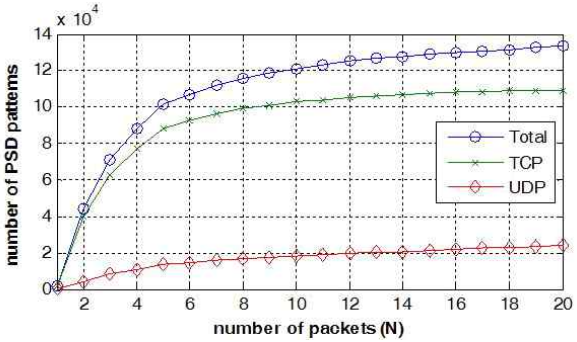
3.2 PSD 충돌에 의한 분류 한계점

본 절에서는 PSD 기반 트래픽 분류 방법의 분석률 또는 정확도를 저하시키는 PSD 충돌에 대해 기술한다. 정답지 분석을 통해 PSD 충돌을 발생시키는 패턴의 양과 트래픽 양을 살펴봄으로써 PSD만으로 분류 가능한 트래픽의 양을 조사한다.

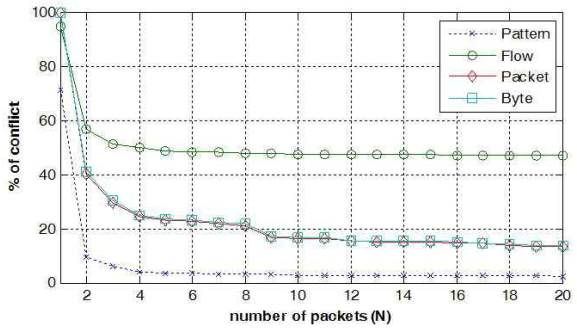
PSD 충돌이란 하나의 PSD 패턴을 하나의 응용이 아닌 여러 개의 응용이 사용하는 것을 말한다. PSD 충돌 플로우는 PSD 만으로는 하나의 응용으로 분류할 수 없다. 따라서, PSD 충돌은 PSD 기반 트래픽 분류 방법의 분석률 또는 정확도를 저하시킨다.

(그림 3)은 제안한 페이로드 크기 분포에서 N 값, 즉 데이터 패킷 수에 따른 PSD 패턴의 양을 나타낸다. N의 값이 증가함에 따라 벡터의 차원이 증가하므로 표현할 수 있는 PSD 패턴의 종류는 많아진다. 하지만, 많은 플로우들이 적은 개수의 데이터 패킷을 포함하는 경우(N의 값이 작은 경우)가 많고, 플로우에서 후기에 나타나는 데이터 패킷은 순수한 데이터 전송용으로 흔히 사용되어 일정한 패턴(e.g. 가장 큰 크기의 데이터 패킷들의 연속)을 지닌다. 따라서, N의 값이 증가하더라도 나타나는 패턴의 수는 기하급수적으로 늘지 않는다.

(그림 4)는 데이터 패킷 수의 변화에 따른 충돌 PSD 패턴과 트래픽의 양을 나타낸다. N이 2이상일 때 충돌 PSD 패턴의 양은 적은 반면, 트래픽 양으로 보았을 때 플로우 단위로 약 50% 정도의 트래픽이 충돌 난다. N의 값이 증가하더라도 충돌 나는 트래픽의 양이 급격히 줄지 않으며, 항상 PSD 충돌이 발생한다. N 값은 시그니쳐 생성과 트래픽 분류에 대한 계산 복잡도에 영향을 미친다. N 값이 커질수록 계산 복잡도는 증가하지만, 분류 가능한 트래픽의 양은 많아 진다. 이는 상충(trade off) 관계로 적절한 N 값을 찾는 작업이 필요하다. 실험을 통해 플로우, 패킷, 바이트 단위의 변화률(N 일 때의 충돌 트래픽 양 - (N-1) 일 때의 충돌 트래픽 양, N은 2이상)이 모두 약 1% 가 되는 이전 시점을 N 값으로 정하였으며, 그 값은 5이다.



(그림 3) 데이터 패킷 수에 따른 PSD 패턴의 양



(그림 4) 데이터 패킷 수에 따른 충돌 트래픽의 양

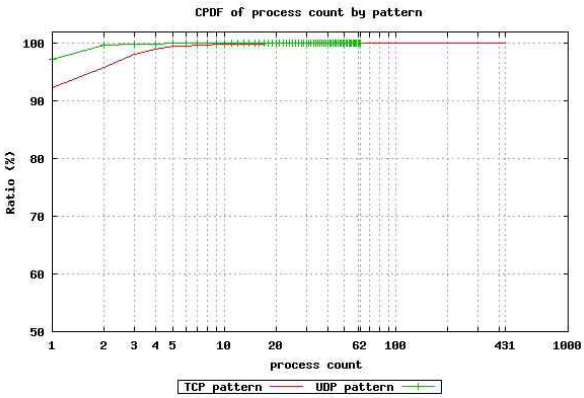
분석을 통해 PSD 충돌을 일으키는 트래픽의 양이 많다는 것을 파악하였다. 이러한 충돌 트래픽은 PSD 기반 트래픽 분류 방법에서 낮은 분석률의 원인이 되거나 잘못 분류하여 정확도를 저하시킬 수 있다. 따라서, 다른 추가적인 분

석을 통해 이러한 충돌 트래픽에 대한 분류 방법이 필요하다. 다음 절에서 이러한 분류 한계점을 극복하기 위한 방안을 제안한다.

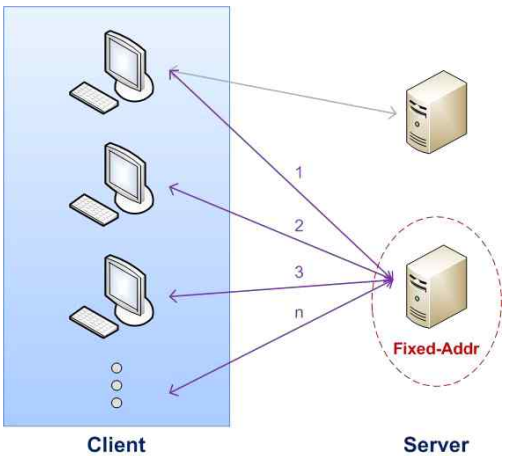
3.3 한계점 극복 방안

PSD 충돌이 발생한 패턴의 트래픽은 추가적인 분석이 있어야 분류가 가능하다. 추가적인 분석 방법을 제시하기에 앞서 얼마나 많은 PSD 패턴이 얼마나 많은 프로세스 사이에서 충돌을 발생시키는 지 살펴본다.

(그림 5)는 충돌이 발생한 응용 수에 대한 PSD 패턴의 누적 비율을 나타낸다. 가로 축은 프로세스의 개수를 세로 축은 PSD 패턴의 비율을 나타낸다. 프로세스의 개수가 1이라는 것은 충돌이 발생한 패턴이 아니라는 뜻이며, 프로세스의 개수가 2라는 것은 충돌을 발생시킨 프로세스가 2개라는 뜻이다. 90% 이상의 패턴은 충돌을 발생시키지 않으며, 충돌을 발생시킨 패턴이라도 할지라도, TCP의 경우 20개 이내, UDP의 경우 5개 이내의 응용에서 충돌을 발생시킨다. 즉, PSD 충돌이 발생하더라도 소수의 응용 중 하나로 결정 가능하다면 많은 응용 중 하나로 결정하는 것보다 트래픽을 쉽고 정확하게 분류할 수 있다.



(그림 5) 충돌 프로세스 수에 따른 PSD 패턴 비율



(그림 6) 고정 IP와port 추출 방법

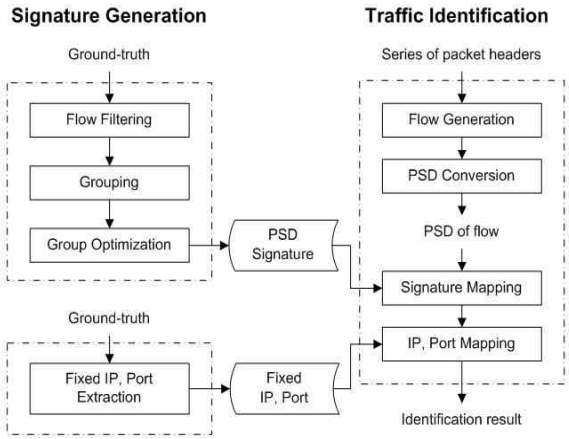
본 논문에서는 IP와 port를 사용하여 PSD 충돌을 해결하고자 한다. 특정 플로우가 충돌 PSD 패턴을 나타내는 경우, 충돌을 발생시키는 응용 사이에서만 IP와 port를 비교하여 플로우를 분류한다. 이 때 사용되는 IP와 port의 리스트는 (그림 6)과 같은 방법을 통해 얻는다.

서버는 서비스를 제공하기 위해 고정된 IP와 port를 가진다. 현재 많은 서버-클라이언트 모델에서 이와 같이 고정된 IP와 port를 가지고 있으며 이를 이용한다면 트래픽 분류가 가능하다[18]. 추출 방법은 하나의 서버가 존재할 때, 2개 이상의 클라이언트가 해당 서버와 데이터 전송을 하는 경우 이 서버의 IP와 port를 테이블에 저장한다. 이렇게 저장된 테이블은 트래픽 분류 시 PSD 충돌이 일어난 플로우가 있을 경우 사용된다.

4. 통계 시그니쳐 생성 및 분류 방법론

본 장에서는 기존 연구[11]에서의 통계 시그니쳐 기반 분류 방법에서 IP와 port를 추가적으로 이용하여 분석률을 향상시킨 방법을 제안한다. 또한, 통계 시그니쳐와 플로우의 거리 측정 방법을 기존의 벡터 거리에서 패킷 별 거리로 변경하여 정확도를 향상시킨다.

(그림 7)은 통계 시그니쳐 생성과 분류 방법에 대한 전체적인 개요이다. 왼편의 시그니쳐 생성 단계는 트래픽 분류에 필요한 응용 프로그램들의 통계 시그니쳐를 생성하는 단계이며, 오른편의 트래픽 분류 단계는 생성된 통계 시그니쳐를 통해 트래픽을 분류하는 단계이다.



(그림 7) 통계 시그니쳐 생성과 분류 방법론의 개요

통계 시그니쳐 생성과 분류 방법은 기존 연구[11]에서 제안한 방법에서 향상된 방법을 적용하기 위해 일부분이 변경되었다. 시그니쳐 생성 단계에서의 변화는 그룹 최적화 단계의 변경, 2장 관련 연구에서 기술한 충돌 처리 단계의 삭제, 그리고 패킷 별 거리 측정을 위한 시그니쳐의 표현 방법의 변경이다. 트래픽 분류 단계에서의 변화는 거리 측정 방법의 변경과 IP, port를 이용한 분류 방법의 추가이다.

4.1 통계 시그니쳐 생성 방법

본 절에서는 변경된 통계 시그니쳐의 생성 방법에 대해 기술한다.

플로우 필터링(전처리) 단계와 그룹핑 단계는 기존 연구[11]의 알고리즘을 따른다. 필터링 단계에서는 시그니쳐 생성의 입력인 정답지 트래픽 중에서 불필요한 트래픽을 제거한다. 그룹핑 단계는 자율 학습(unsupervised learning)으로써 PSD 벡터 간의 거리를 통해 PSD 패턴이 유사한 플로우들을 그룹으로 만들어주는 역할을 한다.

그룹의 최적화 작업은 다음과 같다. 그룹핑 과정 이후 최종 생성된 그룹에 속하지 않는 플로우들을 제거하는 작업은 기존 알고리즘과 동일하다. 변경된 알고리즘에서는 시그니쳐의 정확도를 높이기 위해 플로우의 개수가 적은 그룹을 제거하는 작업이 추가된다. 이는 응용의 특징이 되는 PSD 패턴이라면 자주 등장할 것이라 판단되기 때문이다. 또한, 그룹의 거리 임계값 최소화 작업과 포트 할당 작업을 제거한다. 기존 연구에서는 트래픽 분류 단계에서 시그니쳐가 포트를 포함하는 경우 시그니쳐와 플로우의 PSD 거리뿐만 아니라 포트 번호까지 일치해야 해당 응용으로 분류하였다. 이는 정확도를 높여주는 효과를 가져오지만, 본 논문에서는 시그니쳐와 플로우의 PSD 거리를 통해 분류하지 못한 플로우에 대해서만 포트와 IP를 이용한다.

분류 단계에서 각 패킷 별 거리 측정을 위해 기존 시그니쳐에서 다음과 같이 변경한다. 그룹핑 작업을 통해 하나의 그룹에는 여러 개의 플로우가 포함되어 있다. 기존에는 분류를 위해 그룹의 중심값과 중심과 가장 거리가 먼 플로우를 기준으로 거리 임계값을 시그니쳐로 사용하였다. 하지만, 본 논문에서는 각 패킷 별 거리 측정을 위해 다음과 같이 각 패킷에 대한 거리 임계값을 추출한다.

$$DT_i = \{DT_i \mid \max(|s_i(g) - s_i(f)|), f \in G, 1 \le i \le N\}$$

DTi는 각 패킷에 대한 거리 임계값을 나타내며, i는 플로우에서 데이터 패킷의 순서를 의미한다. 그룹을 G, 그룹의 중심값을 g, 그룹에 속한 플로우를 f라 하고, g와 f의 i번째 데이터 패킷의 페이로드 크기를 각각 si(g), si(f)라 표현한다. 각 패킷의 거리 임계값을 의미하는 DTi는 특정 패킷에 대한 그룹의 중심값과 그룹에 속하는 모든 플로우의 차이 중 가장 큰 값으로 결정된다.

생성된 통계 시그니쳐는 하나의 응용 이름과 프로세스 이름, 전송 계층 프로토콜, PSD 벡터, 각 패킷에 대한 거리 임계값을 가지고 있다.

4.2 거리 계산 방법의 변경

기존의 PSD를 사용한 분류 방법[11]에서는 벡터의 거리를 이용하여 트래픽을 분류한다. 기존 분류 방법은 다음과 같다. 먼저 분류하고자 하는 플로우는 통계 시그니쳐의 전송 계층 프로토콜, PSD 벡터의 차원(초기 데이터 패킷의 개수)이 일치하는지 검사한다. 일치할 경우 분류하고자 하는 플로우를 PSD 벡터로 변경하고 시그니쳐와 거리를 측정하여 거리 임계값 이하일 경우 해당 응용으로 분류한다. 이 때의 거리 측정에 사용된 거리는 여러 가지 벡터 간의 거리 측정 방법 중 맨해튼 거리(Manhattan distance)를 이용한다.

본 논문에서는 좀 더 정확한 트래픽 분류를 위해 각 패킷 별 거리를 이용한 분류 방법을 제안한다. 시그니쳐에는 N개의 데이터 패킷에 대한 거리 임계값이 표현되어 있다. 해당 시그니쳐의 응용으로 분류 되기 위해서는 시그니쳐의 i번째 데이터 패킷의 페이로드 크기와 분류하고자 하는 플로우의 i번째 데이터 패킷의 페이로드 크기의 차이가 DTi보다 같거나 작아야 한다.

PSD 시그니쳐 생성 시 그룹핑 되는 그룹들을 살펴보면 특정 위치의 데이터 패킷이 고정된 크기를 가지는 경우가 많다. <표 2>는 최종 생성된 그룹 중 패킷 별 고정 크기를 가진 그룹의 양을 나타낸다.

〈표 2〉 패킷 별 고정 크기를 가진 그룹의 양

	Total	Non-fixed	Fixed
Group	9,414(100%)	1,711(18.2%)	7,703(81.8%)

<표 2>에서 알 수 있듯이, 생성된 그룹 중 80% 이상의 그룹이 최소한 하나 이상의 고정된 페이로드 크기의 패킷을 보유하고 있다. 벡터의 거리를 이용한 방법보다 각 패킷에 대한 거리를 측정하는 방법을 적용할 경우 고정된 페이로드 크기를 갖는 패킷에 대한 거리 측정은 하지 않고, 페이로드 크기 범위를 갖는 패킷에 대해서만 거리를 측정하므로 계산 속도가 빨라지며, 분류 정확도도 높아진다.

〈표 3〉 분류 정확도(벡터 거리 vs. 패킷 별 거리)

Distance Method	Flow	Packet	Byte
Vector Distance	99.39%	99.12%	99.43%
Packet Distance	99.52%	99.31%	99.49%
Diff.	+0.13%	+0.19%	+0.06%

벡터 거리 측정에 비해 패킷 별 거리 측정의 우수성을 검증하기 위해 3시간 동안의 트래픽에 대해 실험하였다. <표 3>은 같은 트래픽에 대해서 벡터 거리를 이용한 방법과 패킷 별 거리를 이용한 방법의 트래픽 분류 정확도를 나타낸 표이다.

실험을 통해 큰 정확도 향상을 보이진 않았지만, 플로우 단위로 약 0.13%의 정확도 향상을 확인할 수 있었다.

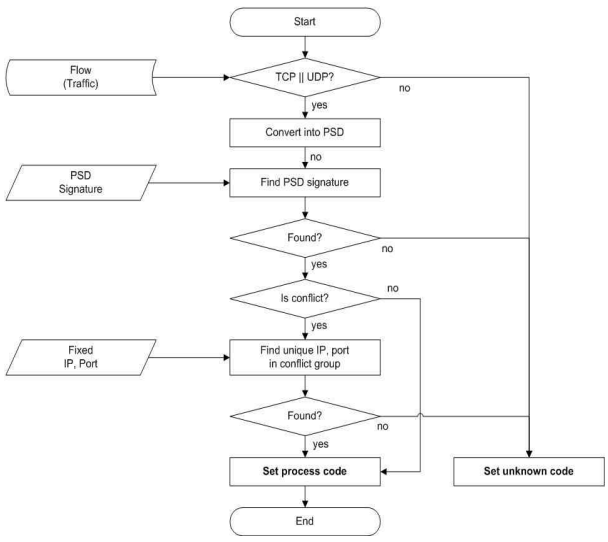
4.3 트래픽 분류 방법

본 절에서는 4.1절의 통계 시그니쳐 생성 방법을 통해 생성된 통계 시그니쳐를 통해 트래픽을 분류하는 방법에 대해 설명한다.

(그림 8)은 통계 시그니쳐를 이용한 트래픽 분류 알고리즘의 순서도를 나타내고 있다. 분류하고자 하는 트래픽, 즉 플로우가 입력으로 들어오면 TCP 인지 UDP 인지 검사하게 된다. 이는 응용 프로그램이 사용하는 전송 계층 프로토콜은 현재 오직 TCP와 UDP 만 사용되므로, 그렇지 않은 경우에는 unknown 코드를 삽입한다. 그런 다음, 플로우를 PSD 표현으로 변경하게 되고, PSD 시그니쳐들 중에서 거리 임계값에 해당되는 모든 시그니쳐를 찾는다. 찾아진 시그니쳐가 하나인 경우, 충돌이 발생하지 않은 것이므로 해당 플로우를 해당 응용으로 분류한다. 만약 찾아진 시그니쳐가 2개 이상의 응용에 해당하는 시그니쳐인 경우에는 PSD 충돌이라고 판단되며, IP-port 테이블에서 응용들의 IP, port를 검사한다.

IP-port 테이블은 각 응용들에 대한 고정 IP, port들이 저장되어 있으며, 검사 방법은 PSD 충돌을 발생시킨 플로우가 어떤 응용들 사이에서 충돌했는지 파악하고, 테이블에서 이들 응용의 IP, port 중 일치하는 IP, port를 찾는다. 찾아진 경우 해당 응용으로 분류하고, 찾지 못하거나 IP, port가 두 개 이상의 응용과 겹치는 경우 분류하지 않는다.

PSD 충돌에 대한 IP, port 검사가 기존 알고리즘과 다른 추가적인 방법이며, 여기서 중요한 점은 모든 응용에 대한 IP, port 검사를 행하는 것이 아니라 충돌을 발생시킨 응용의 IP, port에 대해서만 검사를 실시한다.



(그림 8) 트래픽 분류 알고리즘의 순서도

5. 실험 및 결과 분석

본 장에서는 4장에서 제안된 통계 시그니쳐 기반 응용 트래픽 분류 방법에 대한 실험과 결과 분석에 대한 내용을 기술한다.

총 3일의 학내 망 트래픽을 대상으로 기존 통계 시그니쳐 기반 분류 방법[11]과 본 논문에서는 제안하는 향상된 분류 방법의 결과를 비교한다.

〈표 4〉 분석률

Date	Flow		Packet		Byte	
	prev.	cur.	prev.	cur.	prev.	cur.
2010-10-12	14.59%	23.20%	29.05%	33.37%	32.93%	35.63%
2010-10-17	22.12%	31.43%	28.80%	32.92%	30.19%	33.63%
2010-11-24	19.77%	33.99%	23.22%	27.59%	25.44%	27.62%

〈표 5〉 정확도

Date	Flow		Packet		Byte	
	prev.	cur.	prev.	cur.	prev.	cur.
2010-10-12	99.57%	99.57%	99.28%	99.76%	99.33%	99.81%
2010-10-17	98.80%	98.82%	99.44%	99.73%	99.49%	99.78%
2010-11-24	99.24%	99.38%	99.54%	99.78%	99.59%	99.84%

<표 4>는 각각 기존 알고리즘과 본 논문에서 제안하는 향상된 알고리즘의 분석률을 비교한 표이다. 본 논문에서 제안하는 IP, port 를 통한 PSD 충돌 해결 방법으로 인해 기존 방법론에 비해 플로우 단위로 약 10%의 분석률 향상을 보인다. <표 5>는 분류 정확도는 나타낸 표로써, 제안한 방법론이 기존 방법론에 비해 3일 모두 비슷하거나 향상된 정확도를 나타낸다. 즉, 본 논문에서 제안한 방법론은 기존 방법의 높은 정확도를 유지하면서 기존 방법론의 문제점인 낮은 분석률을 보완하였다고 판단할 수 있다.

〈표 6〉 주요 응용에 대한 precision, recall

Application	Type	Precision	Recall
Torrent	P2P	100.00%	80.72%
Fileguri	P2P	99.91%	94.30%
Windows	System	100%	87.90%
Ms office outlook	Mail	98.91%	91.07%
Internet explorer	Web	100.00%	11.78%
Nateon	Instant Messenger	94.82%	25.17%
Mfile	Web disk	45.83%	1.32%
Kdisk	Web disk	1.06%	0.11%

<표 6>은 학내 망에서 발생하는 주요 응용들의 precision과 recall을 나타내고 있다. Precision과 recall은 각 응용의 분류의 정확도를 나타내는 것으로써 <표 7>과 아래 수식으로 표현된다.

Precision은 'A 그룹으로 분류된 원소 중 실제 A 그룹에 속하는 원소 개수 / A 그룹으로 분류된 전체 원소 개수' 이며, recall은 'A 그룹으로 분류된 원소 중 실제 A 그룹에 속한 개수 / 분류 전 실제 데이터 set에서 A 그룹의 원소 개수' 이다.

Torrent, Fileguri, Windows, Ms office outlook은 모두 98% 이상의 precision과 80% 이상의 recall 값을 가지고 있다. 이는 분류한 트래픽에 대해서 정확하게 분류하였고, 응

〈표 7〉 Confusion matrix

	Actual Positive	Actual Negative
Predicted Positive	TP (True Positive)	FP (False Positive)
Predicted Negative	FN (False Negative)	TN (True Negative)

$$\text{Precision} = \frac{TP}{TP+FP}, \quad \text{Recall} = \frac{TP}{TP+FN}$$

용의 모든 트래픽 중 대부분의 트래픽을 분석한 것이다. Internet explorer와 Nateon은 정확하게 트래픽을 분류하지만 응용의 트래픽 중 많은 트래픽을 분석하지 못한 것이다. 그리고 웹 디스크인 Mfile과 Kdisk는 낮은 precision과 recall을 나타낸다. 웹 디스크는 동일한 엔진과 동일한 데이터 센터를 여러 업체에서 사용을 하여 본 트래픽 분류 방법으로 트래픽을 정확하게 분류하기 어렵다. 웹 디스크에 대한 분류 기준과 정확한 분류 방법에 대한 연구가 필요하다.

6. 결론 및 향후 과제

본 논문에서는 기존 분류 방법론[11]의 낮은 분석률을 해결하기 위해 정답지 분석을 통한 페이로드 크기 분포의 분류 한계점과 극복 방안을 제시하였다. 낮은 분석률은 PSD 충돌에 의해 분류하지 못한 트래픽 때문이며, 이러한 충돌을 해결하기 위해 IP와 port를 이용하였다. 실험을 통해 본 방법론의 향상된 분석률을 검증하였으며, 또한, 거리 측정 방법의 개선을 통해 정확도 향상을 얻을 수 있었다.

향후 연구에서는 페이로드 크기 분포뿐만 아니라 패킷 간 시간 간격 등과 같은 다른 통계적 특징을 이용한 트래픽 분류 방법에 관한 연구를 계획 중이다.

참 고 문 헌

[1] 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.

[2] 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.

[3] Rentao Gu, Minhuo Hong, Hongxiang Wang, and Yuefeng Ji, "Fast Traffic Classification in High Speed Networks," Proc. of the Asia-Pacific Network Operations and Management Symposium (APNOMS) 2008, LNCS 5297, Beijing, China, Oct., 22-24, 2008, pp.429 - 432.

[4] Ying-Dar Lina, Chun-Nan Lua, Yuan-Cheng Laib, Wei-Hao Penga and Po-Ching Lina, "Application classification using packet size distribution and port association" Proc. of the Journal of Network and Computer Applications, In Press, Corrected Proof, Available online, March, 20. 2009.

[5] Huifang Feng and Yantai Shu, "Statistical Analysis of Packet Interarrival Times in Wireless LAN," Proc. of the Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007. International Conference, Shanghai, China, Sept. 21-25, 2007, pp.1888-1891.

[6] Thuy T.T. Nguyen and Grenville Armitage, "A Survey of Techniques for Internet Traffic Classification using Machine Learning," IEEE Communications Surveys and Tutorials, to appear, 2008.

[7] L.Bernaille, R. Teixeira, and K. Salamatian, "Early Application Identification," In: CoNext 2006. Conference on Future Networking Technologies., 2006.

[8] Young-Tae Han and Hong-Shik Park, "Game Traffic Classification Using Statistical Characteristics at the Transport Layer," ETRI Journal, Vol.32, No.1, Feb., 2010, pp.22-32.

[9] Gerhard Munz, Hui Dai, Lothar Braun, and Georg Carle, "TCP Traffic Classification Using Markov Models," In Proc. of Traffic Monitoring and Analysis Workshop (TMA) 2010, Zurich, Switzerland, April, 2010.

[10] Valentin Carela-Espanol, Pere Barlet-Ros, Marc Sole-Simo, Alberto Dainotti, Walter de Donato, and Antonio Pescape, "K-dimensional trees for continuous traffic classification," In Proc. of Traffic Monitoring and Analysis Workshop (TMA) 2010, Zurich, Switzerland, April, 2010.

[11] 박진완, 윤성호, 박준상, 이상우, 김명섭, "통계 시그니쳐 기반의 응용 트래픽 분류", 통신학회논문지 Vol.34 No.11, , Nov., 2009, pp.1234-1244.

[12] 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, April. 7-11, 2008, 160-167.

[13] Ying-Dar Lina, Chun-Nan Lua, Yuan-Cheng Laib, Wei-Hao Penga and Po-Ching Lina, "Application classification using packet size distribution and port association" Proc. of the Journal of Network and Computer Applications, In Press, Corrected Proof, Available online, March, 20. 2009.

[14] Huifang Feng, Yantai Shu, "Statistical Analysis of Packet Interarrival Times in Wireless" Proc. of the Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007. International Conference, Shanghai, China, Sept., 21-25, 2007, pp.1888-1891.

[15] Jacobus van der Merwe, Ramon Caceres, Yang-hua Chu, and Cormac Sreenan "mmdump - A Tool for Monitoring Internet Multimedia Traffic," ACM Computer Communication Review, 30(4), October, 2000.

[16] 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.
[17] Y.J. Won, B.C. Park, H.T. Ju, M.S. Kim, and J. W. Hong. "A hybrid approach for accurate application traffic identification," In IEEE/IFIP E2EMON, April, 2006.
[18] Sung-Ho Yoon, Jin-Wan Park, Young-Seok Oh, Jun-Sang Park, and Myung-Sup Kim, "Internet Application Traffic Classification Using Fixed IP-port," Proc. of the Asia-Pacific Network Operations and Management Symposium (APNOMS) 2009, LNCS5787, Jeju, Korea, Sep., 23-25, 2009, pp.21-30.

박 진 완

e-mail : jinwan_park@korea.ac.kr
2009년 고려대학교 컴퓨터정보학과(학사)
2011년 고려대학교 컴퓨터정보학과(석사)
2011년~현 재 LG전자 근무
관심분야 : 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

김 명 섭

e-mail : tmskim@korea.ac.kr
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~현 재 고려대학교 컴퓨터정보학과 부교수
관심분야 : 네트워크 관리 및 보안, 트래픽 모니터링 및 분석, 멀티미디어 네트워크