http://dx.doi.org/10.3745/KTCCS.2013.2.8.335

Performance Improvement of the Statistical Information based Traffic Identification System

An Hyun Min† · Ham Jae Hyun†† · Kim Myung Sup†††

ABSTRACT

Nowadays, the traffic type and behavior are extremely diverse due to the growth of network speed and the appearance of various services on Internet. For efficient network operation and management, the importance of application-level traffic identification is more and more increasing in the area of traffic analysis. In recent years traffic identification methodology using statistical features of traffic flow has been broadly studied. However, there are several problems to be considered in the identification methodology base on statistical features of flow to improve the analysis accuracy. In this paper, we recognize these problems by analyzing the ground-truth traffic and propose the solution of these problems. The four problems considered in this paper are the distance measurement of features, the selection of the representative value of features, the abnormal behavior of TCP sessions, and the weight assignment to the feature. The proposed solutions were verified by showing the performance improvement through experiments in campus network.

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

통계 정보 기반 트래픽 분석 방법론의 성능 향상

안 현 민† · 함 재 현†† · 김 명 섭†††

요 약

네트워크의 고속화와 다양한 서비스의 등장으로 오늘날의 네트워크 트래픽은 복잡 다양해지고 있다. 효율적인 네트워크 관리를 위해서 QoS, SLA와 같은 정책을 적용하기 위해서는 트래픽 분석 중에서도 응용 트래픽 분류의 중요성이 크다. 현재까지 트래픽 분류에 관한 연구가 활발히 진행되어 왔는데 최근에는 플로우의 통계 정보를 이용한 트래픽 분류 방법론이 많이 연구되고 있다. 하지만 플로우의 통계 정보를 이용한 트래픽 분류 방법론에는 필히 고려해야 할 여러 문제점이 있다. 본 논문에서는 정답지 트래픽 분석을 통해 통계 정보 기반 트래픽 분석 방법론의 해결해야 하는 문제점들을 분석하고 그 해결방안에 대해 제안한다. 통계 정보 기반 트래픽 분석 방법론에서 필히 해결해야 할 문제점은 총 네 가지로 Feature들의 거리 측정 방법과 대표값 추출 방법, TCP 세션의 이상동작, 그리고 패킷 별 가중치이다. 제안하는 방법은 선정한 통계 시그니쳐 기반 트래픽 분석 시스템을 이용한 학내 망에서의 실험을 통해 그 성능을 검증한다.

키워드 : 트래픽 분류, 트래픽 분석, 통계 시그니쳐, 응용 트래픽

1. 서 론

네트워크의 고속화와 다양한 서비스의 등장으로 오늘날의 네트워크 트래픽은 복잡 다양해지고 있다. 이러한 상황 속에서 효율적인 네트워크 관리를 위해 트래픽 분석의 중요성은 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. 이러한 이유로 예전부터 트래픽 분류에 관한 연구는 활발히 진행되어 왔다. 많은 분류 방법이 개발되었지만, 최근에는 플로우의 통계 정보를 이용한 트래픽 분류 방법[3, 4, 5]이 많이 연구되고 있다.

플로우의 통계 정보를 이용한 분류 방법은 패킷 크기, 패킷 간의 시간 간격, 윈도우 크기 등 플로우를 구성하는 패킷들로부터 얻어지는 다양한 통계적 특징을 이용하여 머신 러닝의 특정 알고리즘들을 사용하여 트래픽을 분류하는 방법이 주로 제안되어 왔다[6]. 또한, 특정 통계적 정보를 이용

※ 이 논문은 2012년 정부(교육과학기술부)의 재원으로 한국연구재단(2012R1A1A2007483) 및 2013년도 정부(미래창조과학부)의 재원으로 한국연구재단-차세대정보.컴퓨팅기술개발사업(2010-0020728)의 지원을 받아 수행된 연구임.
† 준 회 원 : 고려대학교 컴퓨터정보학과 석사과정
†† 정 회 원 : 고려대학교 컴퓨터정보학과 박사과정, 국방과학연구소 선임연구원
††† 종신회원 : 고려대학교 컴퓨터정보학과 부교수
논문접수 : 2013년 3월 26일
수 정 일 : 1차 2013년 5월 20일
심사완료 : 2013년 7월 4일
* Corresponding Author : Kim Myung Sup(tmskim@korea.ac.kr)

하여 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.

본 논문에서는 기존 통계 정보 기반 트래픽 분류 방법의 분류 한계점들을 파악하고, 그 한계점을 극복하는 방법을 제시한다. 본 논문에서 다루는 한계점으로는 Feature 사이의 거리 측정 방법, Feature에서 대표값을 추출하는 방법, 트래픽 수집 지점에서의 TCP 세션의 이상동작, 그리고 패킷 별 일정한 가중치가 있다.

성능 검증을 위해 실험에 사용할 통계 시그니쳐 기반 트래픽 분류 시스템을 선정하고 해당 시스템에 제안하는 방법들의 적용 전후의 분석률 및 정확도를 분석하여 그 성능을 검증한다. 제안하는 방법은 학내 망에 분석 시스템으로 구현하고 검증을 통해 실효성을 증명한다.

본 논문은 다음과 같이 구성된다. 서론에 이어 2장에서는 관련 연구에 대해 간략히 설명하고 3장에서는 통계 정보 기반 트래픽 분석 방법의 성능 향상을 위해 고려해야 할 문제들을 분석한다. 4장에서는 3장에서 다룬 문제들의 해결 방안에 대해 설명하며 5장에서는 제안하는 방법의 우수성을 검증하기 위해 소규모 네트워크에서 수집한 트래픽을 이용한 실험 내용과 분석 결과를 기술한다. 마지막으로 6장에서는 결론 및 향후 연구에 대해 기술한다.

2. 관련 연구

본 논문에서 제안하는 방법의 대상은 통계 시그니쳐 기반 분석 방법, 머신 러닝 기반의 분석 방법이다. 통계 시그니쳐 기반 분석 방법은 플로우의 통계정보들을 이용해 응용 별 시그니쳐를 추출하고 이를 통해 트래픽을 응용 별로 분류하는 것이다. 플로우의 통계 정보로는 패킷의 헤더 정보(패킷 크기, 윈도우 크기 등)와 캡쳐 정보(캡쳐 시간, 캡쳐 순서 등)가 있다. 머신러닝 기반의 분석 방법은 응용 트래픽의 특징이 될 수 있는 항목(포트 번호, 플로우 duration, 패킷 간 시간 간격, 패킷 크기 등)들을 머신러닝의 classification clustering기법을 이용하여 트래픽을 분류하는 방법이다. 두 방법은 패킷 크기, 전송 순서 및 방향 등을 사용하므로 본 논문에서 다루는 문제점들을 안고 있다. 본 장에서는 관련 연구로 실험을 위해 선정한 통계 시그니쳐 기반 트래픽 분석 방법[11]을 간략히 설명한다.

[11]에서는 여러 가지 통계적 특징 중에서 페이로드 크기 분포(Payload Size Distribution (PSD))를 플로우 단위로 벡터화 하여 시그니쳐로 생성하고 이를 통해 트래픽을 분류하는 방법을 제안하였다. [11]에서 사용하는 PSD란 양방향 플로우에서 첫 N개 데이터 패킷의 페이로드 크기와 방향을 의미한다. 이 때, 페이로드를 포함하지 않는 TCP 컨트롤 패킷 (SYN, RST, FIN, etc.) 등은 제외한다. 페이로드 크기는 패킷의 페이로드 크기만을 의미한다. 방향은 양수와 음수로 표현하는데 TCP의 경우 양수는 클라이언트에서 서버로 향하는 패킷, 음수는 서버에서 클라이언트로 향하는 패킷을 의미한다. UDP는 서버/클라이언트의 구분이 명확하지 않으므로 첫 패킷을 양수로 표현하고 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. Fig. 1과 같은 플로우를 PSD 벡터로 표현하면 PSD = {+20, -30, +20, +25, -15} 이다.

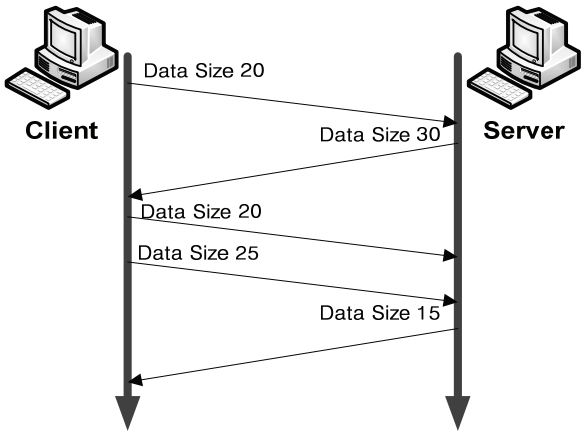

Fig. 1. The bi-directional flow

입력된 플로우를 PSD 벡터로 표현한 뒤 벡터 간 거리가 가까운 플로우들을 그룹핑하고, 그룹 당 하나의 시그니쳐를 추출한다. 시그니쳐는 응용 이름, 전송계층 프로토콜, 각 패킷 별 크기 및 패킷 별 크기 임계값을 가진다. 그룹내 모든 플로우의 목적지 포트가 동일하다면 그 역시 시그니쳐에 포함된다. 시그니쳐의 포함 범위는 2차원에서 직사각형을 이룬다. 직사각형의 중심은 시그니쳐의 패킷 별 크기의 벡터이며 가로, 세로 길이는 패킷 별 임계값을 의미한다.

[11]은 플로우의 PSD벡터를 이용하여 플로우를 그룹핑하고, 각 그룹에서 추출한 시그니쳐를 이용하여 트래픽을 분류하는 방법이다. [11]의 트래픽 분류 방법은 시그니쳐의 포함 범위와 분석 대상 플로우의 PSD 벡터의 비교를 통해 이루어진다. 분석 대상 플로우의 PSD 벡터가 시그니쳐의 포함범위에 존재할 때 분석 대상 플로우를 해당 시그니쳐의 응용으로 분류한다. Fig. 2는 [11]의 트래픽 분류 방법의 예시로 세 개의 플로우 a, b, c에 대한 분류 결과를 보여준다. 플로우 a는 응용 A의 시그니쳐 중 하나에 의해 분류되는 경우로서, PSD 벡터가 응용 A 시그니쳐의 포함 범위에 속하므로 응용 A로 분류한다. 플로우 b는 어떠한 시그니쳐에도 속하지 않는 경우로서, 미확인 (unknown) 트래픽으로 분류한다. 플로우 c는 같은 응용 내에서 두 개 이상의 시그니쳐에 속하는 경우이며, 예시에서는 응용 B로 분류된다.

PSD 기반의 통계 시그니쳐는 응용별 시그니쳐의 충돌이 존재한다. 포함 범위가 겹치는 두 개 이상의 시그니쳐의 응용이 서로 다를 때 이를 충돌이라고 정의한다. 분류 시 충돌 영역에 속하는 플로우는 2개 이상의 응용에 의해 분류가능하므로, 이 중 하나의 응용으로 분류할 경우 잘못 분류할 가능성이 있기 때문에 이는 미확인 (unknown) 트래픽으

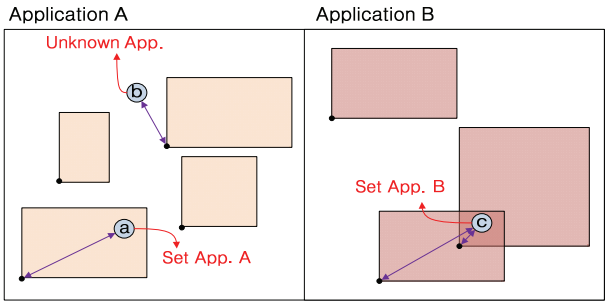

Fig. 2. The example of traffic classification based on the statistic signature

로 분류한다. 본 논문에서는 해당 트래픽 분석 시스템에 제안하는 방법들을 적용하여 통계 정보 기반 트래픽 분석 방법의 성능 향상 방안을 제시한다.

3. 성능 향상을 위한 필수 고려사항

본 장에서는 통계 정보 기반 트래픽 분석 방법에서 고려해야 하는 문제점들에 대해 기술한다.

3.1 거리 계산법

플로우 통계 정보를 이용한 트래픽 분석 방법 중 패킷 크기, 전송 순서 등을 사용하는 방법은 해당 Feature를 N차원으로 표현한다. N의 값이 1이 아닌 경우 N차원 벡터와 N차원 벡터의 거리를 계산하는 방법이 필요하다. 대다수의 경우 N차원 공간에서 두 점 사이의 거리는 Euclidean distance를 사용해 측정한다. Euclidean distance 는 두 점 사이의 최단 거리를 측정하는 방법이다. City-Block distance는 두 점 사이에 장애물이 있을 경우 장애물을 피해 가는 최단 거리를 측정하는 방법이다.

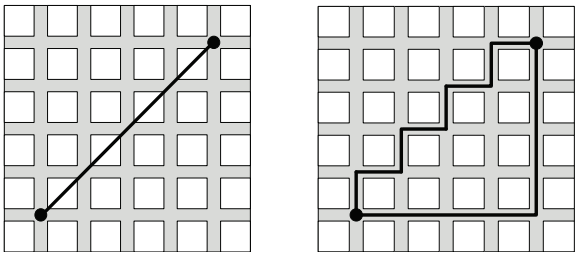

Fig 1A. Euclidean distance　　Fig 1B. City-block distance

Fig. 3. The methods to measure the distance between two points

두 방법은 N차원에서 두 점 사이의 거리를 하나의 값으로 나타내는 방법이다. 이와는 달리 N차원의 두 점 사이의 거리를 N개의 값으로 나타내는 방법으로 각 차원 별로 거리를 구하는 방법이 있다. 즉, 플로우를 N차원 벡터화 하여 벡터 단위의 거리를 측정하는 방법이 아닌, 플로우의 패킷 별 거리를 측정하는 방법이다.

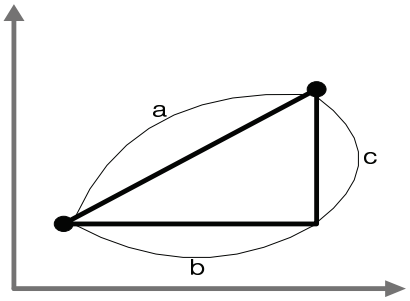

Fig. 4. The distance between two points

Fig. 4와 같이 2차원에 두 점이 있을 때, 두 점의 Euclidean distance는 a 이며 City-Block distance는 $b+c$ 이다. 마지막 패킷 별 거리 측정법은 (b, c) 두 개의 값을 가진다. 세 개의 거리 측정법은 계산속도 및 복잡도가 다르다. 따라서 거리 계산법을 고려해야 한다.

3.2 대표 값 산출

통계 정보 기반 트래픽 분석 방법에서 트래픽을 분석하기 위해서는 Feature를 표현하는 대표 값을 산출해야 한다. 대량의 데이터를 이용하는 해당 방법의 특성 상 모든 Feature가 같은 값을 가질 수 없기 때문에 변위가 존재하며, 따라서 대표 값을 산출하는 방법이 필요하다. 가장 많이 쓰이는 대표 값은 평균이며, 자주 쓰이는 대표 값으로는 최소/최대 값, 중앙값, 최빈값, 그리고 분산과 표준편차가 있다.

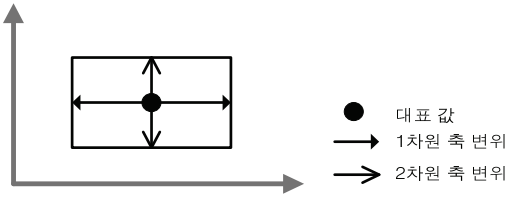

Fig. 5. The representative and variation using median value

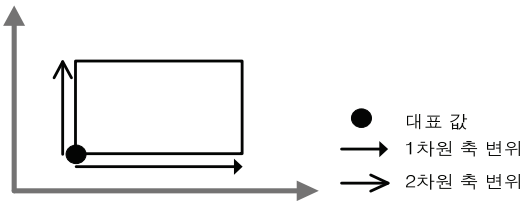

Fig. 6. The representative and variation using minimum value

Fig. 5와 6은 각각 중앙값과 최소값을 대표값으로 하였을 때의 대표값과 변위를 2차원 형태로 표현한 것이다. 포함 범위는 같으나 표현 형태가 달라짐을 볼 수 있다. 대표 값 산출 방법에 따라 변위를 계산하는 방법도 달라지며, 트래픽 분석을 위한 계산 방법도 달라지기 때문에 Feature를 표현하는 대표값을 산출하는 방법에 대해 고려하여야 한다.

3.3 TCP 세션의 이상동작

TCP세션에는 트래픽을 수집하는 지점과 종점 호스트에서 패킷의 순서가 다르게 나타나는 이상동작이 있다. 패킷 Retransmission과 Out-of-order가 그것이다. 가장 많이 쓰이는 전송 계층 프로토콜 중 하나인 TCP 트래픽을 수집하는 과정에서 이러한 문제가 발생한다면 통계 정보 기반 트래픽 분류 방법은 신뢰할 수 없는 분석 결과를 낼 것이다[12].

TCP/IP의 트랜스포트 계층 프로토콜인 TCP는 데이터 전달의 신뢰성을 보장해준다. TCP를 사용하여 종점 호스트 사이에 연결이 설정되면 연속된 바이트 전송을 보장하는 스트림이 제공되는 것이다. 이를 위해 TCP는 Ack(확인 응답)와 Sequence, Acknowledge Number 등을 사용하여 패킷(데이터)의 이상 유무를 확인하고 바로잡는다. Fig. 7과 같은 형태가 정상적인 TCP 세션의 데이터 흐름이다.

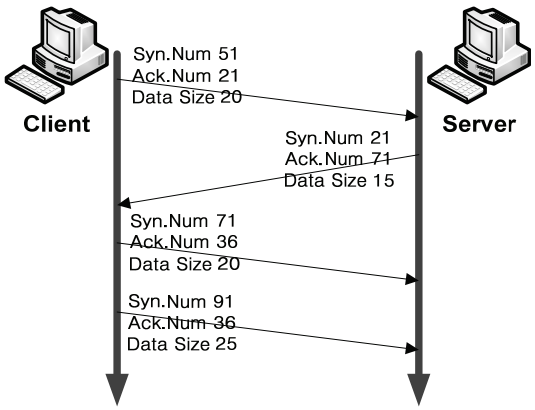

Fig. 7. The normal TCP session

이러한 흐름 중, 고려해야 할 이상동작은 패킷 Retransmission과 패킷 Out-of-order이다. 수신 측에서 에러를 발견하면 에러가 발견된 데이터를 버리고 Ack를 하지 않음으로써 송신 측에 에러가 났음을 알리거나 최근 정상적으로 수신한 데이터에 대한 Ack를 반복하여 보냄으로써 송신 측이 Ack 이후의 데이터에서 에러가 발생한 것(또는 데이터그램이 분실된 것)을 알 수 있도록 하여 Fig. 8과 같이 해당 데이터를 Retransmission(재전송)하도록 한다. 이처럼 데이터에서 에러가 발생하면 (또는 데이터그램이 분실되면) 수신 측에서는 해당 데이터를 버리는 등의 동작을 하여 TCP의 바이트 스트림을 보장받는다.

또한 패킷이 여러 가지 이유로 순서가 뒤바뀌는 경우가 있다. 하나의 호스트에서 상대 호스트로 패킷을 전송할 때 패킷이 거치는 라우터의 경로가 달라지는 경우가 있는데 이때 패킷의 순서가 뒤바뀌어 도착할 수 있다. 이를 패킷 Out-of-order라 하며 Fig. 9와 같은 형태를 띈다.

TCP는 1 번부터 5번까지의 패킷이 순서대로 전송되어야 할 때 2번 패킷보다 3번 패킷이 먼저 도착하였다면 올바른 순서로 온 것이 아님을 파악해 3번 패킷을 버퍼에 쌓아두고 2번 패킷이 도착한 후 순서에 맞춰 2번 패킷을 먼저 저장한 뒤 3번 패킷을 저장한다. 이를 Reordering이라 칭하는데 이

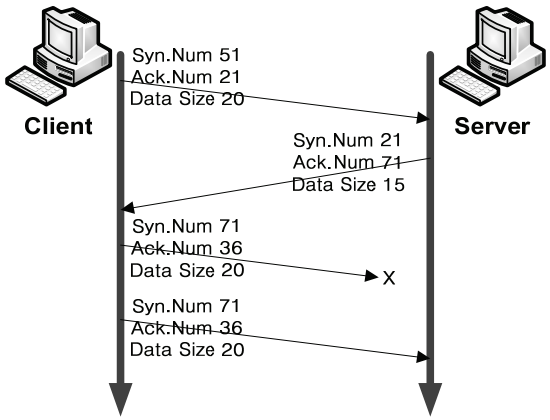

Fig. 8. The packet retransmission in the TCP session

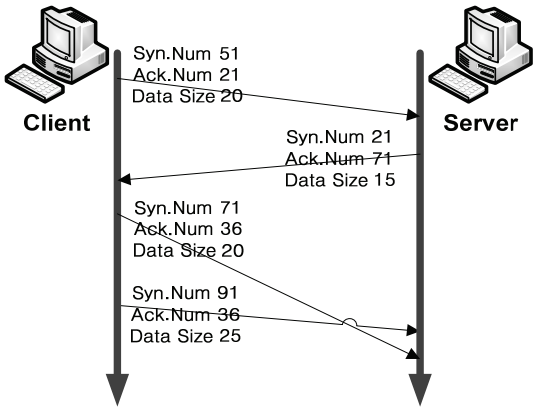

Fig. 9. The packet out-of-order in the TCP session

러한 방법으로 TCP는 바이트 스트림을 보장 받는 것이다. 문제는 TCP의 이러한 바이트 스트림 보장 행위가 종점 호스트에서 일어난다는 것이다. 때문에 트래픽 분석을 위해 중간에서 트래픽을 수집(캡쳐)하는 지점에서는 순서에 맞지 않는 트래픽을 수집하게 된다. [12]에 따르면 수집되는 트래픽의 10%를 넘는 TCP 플로우가 이와 같은 이상동작으로 인해 순서에 맞지 않게 수집된 플로우이다. 이러한 이상동작을 해결하지 않는 것은 통계 정보 기반 트래픽 분류 방법의 한계가 될 수 있다. 따라서 트래픽을 수집하는 지점에서는 수집과 동시에 이러한 이상동작을 해결하여야 한다.

3.4 패킷 별 가중치

본 절에서는 패킷 별 가중치를 이용하는 것의 타당성 입증 실험 결과를 분석한다. 실험은 학내 망에서 수집한 트래픽을 이용하여 진행하였고, 대상으로 하는 방법의 Feature인 패킷 크기, 전송 방향, 전송 순서 중 가중치를 할당하기 적합한 패킷 크기에 가중치를 할당하는 것이 타당한지 검증하기 위해 패킷의 전송 방향과 순서가 같은 플로우들을 그룹핑 하여 각 그룹별로 패킷 분포를 분석한다. 패킷은 데이터의 흐름을 보기 위해 페이로드가 있는 패킷만을 사용하였으며 첫 패킷은 TCP의 경우 3-handshake 패킷 이후의 페이로드가 있는 패킷을 첫 패킷으로 정의한다. 같은 응용에서

발생하고, 패킷 전송 방향과 순서가 같은 플로우를 그룹으로 나누는 것을 그룹핑 조건으로 하였다.

Fig. 10은 그룹들의 패킷 별 분산을 나타낸 분산그래프이다. 그룹핑 결과 총 66개의 그룹이 생성 되었고, 각 그룹 내에서 패킷 별 표준편차를 계산한 후 그 값들을 그래프로 나타낸 것이다.

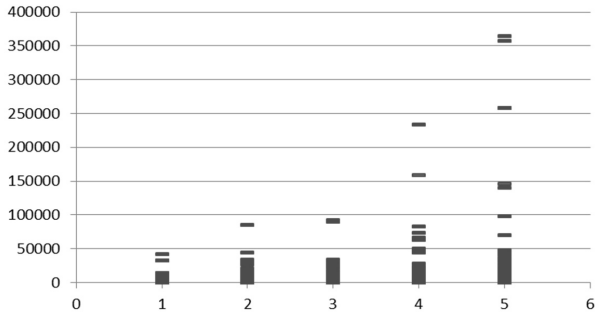

Fig. 10. The variance distribution of packet size in order of packets of each flow that occurred in the same application

그래프를 살펴보면, 1, 2, 3번째 패킷의 분포는 비슷하며 4, 5번째 패킷의 분포 범위가 더 넓다. 그 중 5번째 패킷의 분포 범위가 눈에 띄게 넓다. 즉, 패킷의 전송 순서에 따라 페이로드 크기 변화의 폭과 빈도가 변하는 것이다. 이처럼 전송 순서에 따라 패킷의 크기가 다른 변화를 보인다면 이는 패킷의 크기 및 전송 순서를 이용하여 트래픽을 분류하는 방법에서 모든 패킷에 동일한 가중치를 두는 것 보다는 패킷 순서에 따른 차별적 가중치를 적용하는 것이 분석 효율을 높이는데 더 효과적일 것이다. 따라서 패킷 별로 가중치를 적용하는 방법에 대해 고려해야 한다.

4. 해결 방안

본 장에서는 3장에서 다뤘던 통계 정보 기반 트래픽 분석 방법의 고려 사항들의 해결 방안에 대해 기술한다.

4.1 거리 계산법

3.1절에서 Feature를 N차원으로 표현할 때, N차원 벡터와 N 차원 벡터의 거리를 측정하는 방법으로 두 점 사이의 거리를 하나의 값으로 나타내는, 벡터 단위 거리 측정 방법과 두 점 사이의 거리를 N개의 값으로 나타내는 패킷 별 거리 측정 방법, 크게 두 가지에 대해 기술하였다.

두 점 사이의 거리를 하나의 값으로 나타내는 방법으로는 Euclidean distance와 City-Block distance가 대표적이다. Euclidean distance는 두 점 사이의 최단거리를 구하는 방법이고 City-Block distance는 장애물을 사이에 두고 있는 두 점 사이의 거리를 구하는 방법이다.

$$V(p)=(p_1,p_2,...,p_n) \tag{1}$$

n차원의 점 p의 벡터 $V(p)$는 (1)과 같이 표현한다. $d(V(p),V(q))$가 두 벡터 $V(p)$와 $V(q)$의 거리를 나타낸다면 Euclidean distance는 (2), City-Block distance는 (3)을 이용해 계산한다.

$$d(V(p),V(q))=\sqrt{\sum(p_i-q_i)^2} \tag{2}$$

$$d(V(p),V(q))=\sum|p_i-q_i| \tag{3}$$

대부분의 경우 N 차원의 공간에서 두 벡터간의 거리는 Euclidean distance를 사용한다. 그럼에도 불구하고 통계 정보 기반 트래픽 분석 시스템에서는 City-Block distance를 자주 이용하는데 이는 계산상의 속도 문제 때문이다. 두 개의 거리를 구하는 공식은 서로 비례한다. 하지만 City-Block distance는 Euclidean distance에 비해 계산식이 간단하므로 많은 양의 데이터를 처리해야 하는 트래픽 분석에서는 처리 속도에서 상당한 이점을 가질 수 있다. 특히, 실시간 트래픽 분류에서는 데이터의 처리 속도가 더욱 빨라야 하므로 City-Block distance가 알맞고, 그 때문에 자주 이용된다.

하지만 본 논문에서는 패킷 별 거리 측정법을 제안한다. 패킷 별 거리 측정법은 패킷 각각 1차원의 계산을 하므로 City-Block distance와 계산 속도가 같으며, 하나의 값이 아닌 N개의 값을 가짐으로써 Feture가 최대 N-1개까지 증가한다. 즉, 같은 정보로 더욱 다양하게 Feature를 표현할 수 있다. 또한, [11]에서 실험한 결과에 의하면 유사한 Feature를 가진 플로우들을 그룹핑하여 살펴보았을 때 고정 크기를 띄는 패킷을 가진 그룹의 양이 전체 그룹의 양의 81.8%를 차지한다. 고정 크기를 띄는 패킷은 변위 0 값을 갖고, 이를 Feature로 이용하는 것은 더욱 정확한 트래픽의 분석을 가능하게 한다. 변위가 0인 패킷을 갖는 시그니쳐와 City-Block distance로는 소량의 차이가 나는 플로우도 변위가 0인 패킷의 크기가 다르다면 패킷 별 거리 측정법에서는 분석되지 않기 때문이다. 따라서 통계 정보 기반 트래픽 분석 방법론에서 Feature를 N차원으로 표현할 때, N차원의 두 점 사이의 거리를 측정하는 방법으로는 패킷 별 거리 측정법을 사용해야 한다.

4.2 대표값 산출

3.2에서 기술한 바와 같이 트래픽을 분석하기 위해서는 Feature를 표현하는 대표 값을 산출해야 한다. 통계에서 자주 쓰이는 대표값은 여러 가지가 있는데 본 논문에서는 최소값, 혹은 최대값을 이용하는 것을 제안한다.

대표값으로 평균이나 중앙값을 이용할 경우 변위는 대표값에서 +/- 형태로 존재하고, 최소/최대값을 이용할 경우 한 방향의 변위만을 가진다. 트래픽 분석을 위해선 플로우가 시그니쳐 혹은 클러스터 등에 포함되는지 확인하여야 하며 이를 위해선 포함범위를 계산하여야 한다. 이때, 변위가 +/- 형태로 존재할 경우 범위 계산을 위해선 두 번의 산술 계산

과 두 번의 조건 확인 계산이 필요하다. 반면, 변위가 한 방향만 존재할 경우 한 번의 산술 계산과 한 번의 조건 확인 계산으로 범위 계산이 가능하다. 즉, worst case의 경우 최소/최대값을 대표값으로 하였을 경우 평균 혹은 중앙값을 대표값으로 했을 때에 비해 절반의 계산만으로 분류가 가능하다. 요즘과 같이 인터넷이 발달하여 분 당 트래픽이 대량인 경우 실시간 분석에서 그 성능 차이가 드러날 수밖에 없다. 따라서 본 논문에서는 Feature의 대표값으로 최소/최대값을 사용해서 변위를 이용한 포함 범위 계산을 한 번만 하는 것을 제안한다.

4.3 TCP 세션의 이상 동작

TCP세션의 이상동작으로 인해 종점 호스트에서의 패킷과 트래픽을 수집하는 지점에서의 패킷의 크기, 전송 순서나 그 방향이 서로 다르다면 통계 기반 트래픽 분석 방법론은 신뢰할 수 없는 결과를 나타낸다. 따라서 TCP 세션의 이상동작인 패킷 Retransmission과 Out-of-order를 해결하여야 한다.

두 이상동작은 모두 TCP 패킷의 헤더에 명시되는 Sequence Number(Seq.Num)와 Acknowledge Number (ACK.Num)를 이용하여 해결할 수 있다. 패킷 Retransmission은 연속되어 전송되는 같은 방향의 두 개의 패킷의 Seq.Num이 동일할 경우 먼저 전송된 패킷을 무시하는 방법으로 해결할 수 있다. 데이터에 오류가 발생하여 패킷이 재전송되는 경우에 후에 전송된 패킷이 완전한 패킷이므로 뒤의 패킷이 아닌 먼저 전송된 패킷을 무시해야 한다. 또한, 연속되어 전송되는 같은 방향의 두 개의 패킷의 Seq.Num을 비교하여 후에 전송된 패킷의 Seq.Num이 클 경우, 두 패킷의 순서를 뒤집어주는 방법으로 패킷 Out-of-order를 해결할 수 있다.

통계 기반 트래픽 분석 방법론에서는 이 방법으로 트래픽 수집 단계에서 TCP 세션의 이상 동작을 해결하여야 한다.

4.4 패킷 별 가중치

본 절에서는 3.4절에서 가능성을 살펴 보았던 패킷 별 가중치 적용의 성능을 실험을 통해 검증한다.

패킷 별 가중치 적용 실험에 앞서 패킷 별 가중치를 결정해야 한다. 3.4절의 실험 결과인 각 그룹의 패킷 별 분산을 이용하여 계산하였다. $v(p_i)$는 i번째 패킷의 분산이며 $G_j(v(p_i))$는 그룹 j의 i번째 패킷의 분산이다.

$$Sum_i = \sum_{j=1}^{N} G_j(v(p_i)) \tag{4}$$

먼저 각 그룹의 패킷 별 분산을 더해 (4)와 같이 패킷 별 분산의 합을 구한 후 (5)를 이용해 패킷 별 분산의 합을 모두 더해 전체 분산의 합을 구하고 이를 이용하여 (6)을 계산해 패킷 별 분산 합의 전체 분산 합과의 비율을 계산한다.

$$TotalVariance = \sum Sum_i \tag{5}$$

$$Rate(Sum_i) = Sum_i \times \frac{100}{TotalVariance} \tag{6}$$

패킷 별 분산의 합의 비율을 소수 첫째 자리에서 반올림하여 (7)과 같은 을 얻었고 계산의 편의성을 위해 1의 자리에서 반올림 한 뒤 10으로 나누어 (8)과 같은 를 얻었다.

$$Rate(Sum) = \{7,10,13,26,44\} \tag{7}$$

$$WeightVector = \{1,1,1,3,4\} \tag{8}$$

구하여진 WeightVector를 이용하여 적용 전 후 실험을 하였다. 실험은 기본적으로 앞선 세 개의 문제를 해결한 시스템을 이용하였다. 즉 플로우 그룹핑과 트래픽 분석에서 패킷 별 거리측정법을 이용하였고, 추출하는 시그니쳐의 대표값을 그룹 내 플로우 벡터 각 요소의 최소값으로 계산하였으며 TCP세션의 이상동작을 개선한 트래픽을 사용하였다. 실험에 사용한 트래픽은 Table 1과 같다.

Table 1. The traffic information of collected application traces (ground-truth)

Application	Flow (x 103)	Packet (x 103)	Byte (x 103)
6	142	1,358	580,429

트래픽 분류에 사용되는 시그니쳐를 생성하거나 분류 규칙을 생성하기 위해서는 정답지(ground-truth) 트래픽이 필요하다. 이러한 정답지는 매우 정확해야만 트래픽 분류 결과에 신뢰성을 보장해준다. 본 연구에서는 TMA-에이전트 기반의 정답지 생성 방법[13]을 이용하여 정확한 정답지를 학내 망에서 10일의 기간동안 수집하였다. 응용은 총 6가지로 Dropbox, KartRider, NateOn, Skype, Teamviewer, uTorrent 가 그것이다.

Table 2. The completeness and accuracy of traffic classification using the packet weight

구분	Flow		Packet		Byte	
	적용 전	적용 후	적용 전	적용 후	적용 전	적용 후
분석률	76.67%	76.98%	50.90%	50.93%	44.90%	44.91%
정확도	99.80%	99.82%	99.11%	99.20%	99.40%	99.40%

패킷 별 가중치를 적용하였을 때가 적용하지 않았을 때보다 적게나마 모든 면에서 좋은 결과를 나타냈다. 패킷 별 가중치를 적용함으로써 플로우를 그룹핑 하는 거리 기준 값을 더 크게 설정하여도 높은 정확도를 유지할 수 있었고 이로 인해 분석률의 증가를 보였다.

5. 실험 및 결과 분석

본 장에서는 논문에서 제안한 네 가지 문제의 해결 방안에 대한 실험과 결과 분석에 대한 내용을 기술한다.

총 10일 간 학내 망에서 수집한 트래픽을 대상으로 통계 정보 기반 트래픽 분류 방법론을 선정하여 문제점들의 해결 방안 적용 유무에 따른 분석 결과를 비교하였다. 트래픽 정보는 표 1과 같다. 서로 다른 둘 이상의 응용 시그니처가 하나의 플로우를 포함할 때에는 분류하지 않았다.

표 3은 각각 선정한 통계 기반 트래픽 분석 시스템의 기존 알고리즘과 본 논문에서 제안하는 해결 방안들을 적용한 알고리즘의 성능을 비교한 표이다. 분석률 측면에선 적용 전 보다 후가 플로우 단위로 4.86%, 패킷 단위로 0.1%, 바이트 단위로 0.5% 증가하였다. 정확도 측면에서는 적용 전 보다 후가 플로우 단위 0.02%, 패킷 단위 0.16% 증가하였고 바이트 단위로 소수 둘째 자리까지 같은 결과를 보였다.

Table 3. The completeness and accuracy of traffic classification using the proposed approach

구분	Flow		Packet		Byte	
	적용 전	적용 후	적용 전	적용 후	적용 전	적용 후
분석률	72.12%	76.98%	50.19%	50.93%	44.35%	44.91%
정확도	99.80%	99.82%	99.04%	99.20%	99.40%	99.40%

분석 방법론이 같더라도 결과에 영향을 끼치는 여러 문제들을 해결함으로써 더 좋은 성능을 내는 통계 기반 트래픽 분석 시스템이 된 것을 확인하였다.

본 논문에서 제안한 해결방안들은 기존의 시스템의 성능을 향상시킨 결과를 나타내었다.

6. 결론 및 향후 과제

본 논문에서는 패킷 크기 통계 정보 기반 트래픽 분류 방법론에서 고려해야 할 사항들을 분석하고 해결 방안을 제시하였다. 또한 실험을 통해 그 성능을 검증하였다. Feature 사이의 거리를 측정하는 방법으로는 패킷 별 거리 측정법을 사용해야 하며 Feature의 대표값으로는 중앙값, 혹은 평균값보다 최소/최대값을 사용하여야 한다. 또한 TCP 세션의 이상동작은 트래픽 수집 지점에서 해결하여야 하며, 전송 순서에 따라 패킷 별 가중치를 적용하여야 한다. 그럼으로써 더욱 좋은 성능을 가진 시스템이 될 수 있다.

향후 연구에서는 응용 별로 패킷 별 가중치를 추출할 수 있는 방법과 통계 정보 기반 트래픽 분류 방법론에 영향을 끼치는 다른 요소에 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 T Han, Hong S 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] Jin-Wan Park, Myung-Sup Kim, "Performance Improvement of the Statistic Signature based Traffic Identification System", KIPSTC,.18C.4., Aug., 2011, pp.243-250.

[12] Hyun-Min An, Myung-Sup Kim, "A Method to resolve the Limit of Traffic Classification caused by Abnormal TCP Session", KNOM Review, Vol.15, No.1, Dec., 2012, pp.31-39.

[13] 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, pp.160-167.

안 현 민

e-mail : queen26@korea.ac.kr
2012년 고려대학교 컴퓨터정보학과(학사)
2012년～현 재 고려대학교 컴퓨터정보학과 석사과정
관심분야 : 네트워크 관리 및 보안, 트래픽 모니터링 및 분석

함 재 현

e-mail : jaehyun_ham@korea.ac.kr
1999년 동국대학교 컴퓨터공학과(학사)
2001년 포항공과대학교 컴퓨터공학과 (석사)
2001년～현 재 국방과학연구소 선임연구원
2012년～현 재 고려대학교 컴퓨터정보학과 박사과정, 국방과학연구소 선임연구원
관심분야 : 전술통신망 관리, 네트워크 관리, 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