Weighted Payload Size Sequence 기반 트래픽 분류 방법론

안현민, 함재현, 김명섭
고려대학교

{queen26, jaehyun_ham, tmskim}@korea.ac.kr

Weighted Payload Size Sequence based Traffic Identification Method

Hyun-Min An, Jae-Hyun Ham, Myung-Sup Kim
Korea Univ.

요 약

효율적인 네트워크 관리를 위해서는 네트워크에서 발생하는 트래픽에 대한 다양한 분석이 필요하며 QoS, SLA 와 같은 정책을 적용하기 위해서 트래픽 분류의 중요성이 크다. 최근에는 플로우의 통계 정보를 이용한 트래픽 분류 방법론이 많이 연구되고 있다. 본 논문에서는 응용에서 발생한 플로우 내의 패킷 전송 순서 별 가중치를 이용하는 Weighted Payload Size Sequence 기반 트래픽 분류 방법론을 제안한다. 제안하는 방법론은 2-path 플로우 그룹핑을 통해 시그니쳐를 추출한다. 학내 망에서의 실험을 통해 제안하는 방법론의 성능을 검증한다.

Ⅰ. 서 론

네트워크의 효율적 운용과 관리를 위한 응용 레벨의 트래픽의 모니터링과 분석은 네트워크 사용현황 파악과 확장계획 수립 등의 다양한 분야에서 필요성이 커져가고 있다. 이를 위해서는 다양한 종류의 응용 레벨 트래픽을 정확하게 분류할 수 있는 방법과 고속 링크에서 발생하는 대용량의 트래픽을 실시간으로 처리하는 방법이 요구된다. 최근에는 높은 정확도를 가지며 분석 속도도 빠른, 플로우의 통계 정보를 이용한 트래픽 분류 방법론[1,2]이 많이 연구되고 있다.

플로우의 통계 정보를 이용한 분류 방법은 패킷 크기, 패킷 간의 시간 간격, 윈도우 크기 등 플로우를 구성하는 패킷에서 얻어지는 여러 통계적 요소를 머신 러닝의 특정 알고리즘에 적용하여 트래픽을 분류하는 방법이 주로 제안되어 왔다[1]. 특정 통계적 정보를 이용하여 자체적인 알고리즘을 개발한 연구들도 진행되었는데, 그 중 패킷 또는 페이로드 크기 분포를 이용한 분류 방법들[2,3]이 많이 제안되고 높은 정확도를 나타내었다.

본 논문에서는 WPSS(Weighted Payload Size Sequence)기반 트래픽 분류 방법론을 제안한다. 패킷의 페이로드 크기와 전송 순서 및 방향을 이용하여 응용 별 통계 시그니쳐를 추출하고 트래픽을 분류하는 방법이다. 이는 패킷의 헤더정보와 캡쳐 정보만을 이용하므로 시그니쳐 추출 및 트래픽 분류 속도가 빠르며 높은 정확도를 나타낸다.

본 논문의 구성은 다음과 같다. 본 장의 서론에 이어, 2 장에서는 관련연구에 대해 기술하고, 3 장에서는 제안하는 방법론에 대해 기술한다. 4 장에서는 제안하는 방법과 Payload Size Sequence 를 사용하는 방법의 트래픽 분류 실험 결과 비교를 통해 그 성능을 검증한다. 마지막으로 5 장에서는 결론 및 향후 연구에 대해 기술한다.

이 논문은 정부(교육과학기술부)의 재원으로 2010 년도 한국연구재단-차세대정보컴퓨팅기술개발사업(20100020728) 및 2012 년도 한국연구재단(2012R1A1A2007483)의 지원을 받아 수행된 연구임.

Ⅱ. 관련연구

본 장에서는 통계 시그니쳐 기반 트래픽 분류 방법론에 대해 간략히 설명한다.

통계 시그니쳐란 플로우의 여러 통계적 요소(패킷 크기, 윈도우 크기, 패킷 캡쳐 시간 및 순서 등)에서 추출한, 다른 응용 프로그램과 구별할 수 있는 응용 프로그램의 고유한 통계적 특징을 의미한다. 통계 시그니쳐 기반 트래픽 분류 방법은 응용 별 통계 시그니쳐를 추출하고 이를 이용하여 트래픽을 분류하는 것이다.

[3]은 본 논문에서 제안하는 방법론과 같은 Feature 를 사용하는 통계 시그니쳐 기반 트래픽 분류 방법론을 제안하였다. [3]에서 제안한 방법론은 PSD (Payload Size Distribution) 벡터를 이용한다. 플로우 내의 패킷의 전송 순서, 방향, 페이로드 크기를 이용하여 벡터로 나타낸 것이다. 전송 방향은 TCP 의 경우 Client 에서 Server 로 전송하는 패킷의 방향을 +라 정의하고 UDP 의 경우 첫 번째 패킷의 전송 방향을 +라 정의한다. 그리고 +를 기준으로 전송 방향이 반대일 경우 -라 정의한다. 페이로드 크기에 전송 방향을 곱한 값이 패킷 전송 순서에 맞는 PSD 벡터의 요소가 된다.

먼저 플로우를 PSD (Payload Size Distribution) 벡터화 한다. 모든 플로우를 PSD 벡터화 한 뒤 플로우의 PSD 벡터와 그룹의 대표벡터 사이의 City-block Distance 가 기준치 내일 경우 해당 그룹으로 플로우를 그룹핑하고, 기준치 내에 그룹이 없을 경우 생성한다. 그룹에 속한 모든 플로우의 각 요소 별 평균이 그룹의 대표벡터의 각 요소가 된다. 그룹핑이 끝나면 그룹을 최적화 한 뒤 그룹의 대표 벡터와 패킷 별 거리 임계치를 시그니쳐로 추출한다. 마지막으로 추출된 시그니쳐를 이용해 트래픽을 분류하는 방법론이다.

Ⅲ. 제안하는 방법

제안하는 방법론은 Weighted Payload Size Sequence (WPSS)를 기반으로 한 트래픽 분류 방법론이다.

WPSS 는 플로우의 Payload Size Sequence 에 패킷 순서 별 가중치를 곱하여 벡터화 한 것이다. 플로우 내 패킷의 전송 순서, 방향 및 페이로드 크기를 그 Feature 로 사용한다.

한 응용 내에서 발생한 플로우 중 패킷의 페이로드 크기를 제외한 나머지 Feature(전송 순서, 방향)가 일치하는 플로우들을 그룹핑 하면 각 그룹 마다 패킷 순서 별로 페이로드의 크기 분포가 다르다. 따라서 본 논문에서는 각 그룹 내 패킷 순서 별 페이로드 크기 분포에 따라 각기 다른 가중치를 추출하고 이를 적용하여 재 그룹핑 한 뒤 시그니쳐를 추출한다. 그림 1 은 전체적인 개요이다.

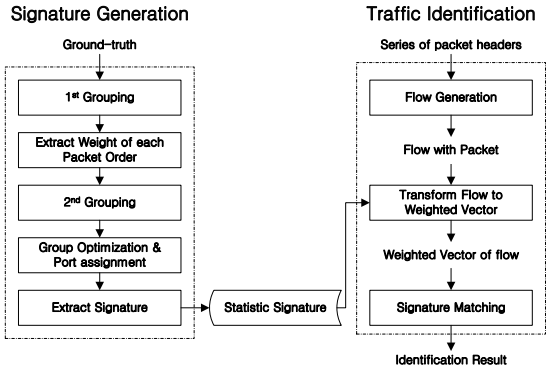

그림 1. 분류 방법론의 개요

왼쪽의 시그니쳐 생성 단계는 트래픽 분류에 필요한 응용 프로그램 별로 통계 시그니쳐를 생성하는 단계이며, 오른편의 실시간 분류 단계는 통계 시그니쳐를 통해 온라인 트래픽을 실시간으로 분석하는 단계이다.

시그니쳐 생성단계에서는 먼저 플로우 내 패킷의 전송 순서와 방향이 일치하는 것을 기준으로 1 차 그룹핑한다. 1 차 그룹핑이 끝나면 생성된 1 차 그룹 내에서 패킷 순서 별 가중치를 추출한다. 가중치는 패킷의 페이로드 크기의 분포를 기준으로 추출하므로 먼저 그룹에 속한 플로우들의 내부 패킷의 전송 순서 별로 페이로드 크기의 분산($variance_i$)을 구한다. 그 후 각 전송 순서 별 분산의 합($TotalVariance$)을 구한다. 마지막으로 (1)과 같이 각 순서 별 가중치($Weight_i$)를 계산한다.

$$Weight_i = TotalVariance / variance_i \quad -(1)$$

모든 1 차 그룹의 패킷 순서 별 가중치가 추출되면 이를 이용하여 2 차 그룹핑을 수행한다. 2 차 그룹핑 단계에선 먼저 각 플로우들을 내부 패킷의 전송 순서와 방향, 그리고 페이로드 크기에 해당하는 가중치를 곱한 값을 이용해 WPSS 벡터화한다. 그 후, 주어지는 각 패킷 별 크기 임계치와 대표 벡터 사이에 플로우 벡터의 모든 요소를 포함할 수 있는 그룹에 그룹핑한다. 그룹의 대표 벡터는 그룹에 속한 모든 플로우들 내부 패킷 페이로드 크기의 평균을 자신의 각 요소로 삼는다. 포함하는 그룹이 없을 경우 새로운 그룹을 생성한다. 대표 벡터는 2 차 그룹핑을 수행하면서 포함되는 플로우가 달라짐에 따라 계속적으로 변하는데 이때 처음에는 그룹에 포함되었으나 그룹핑이 끝나고 그룹을 벗어나는 플로우가 생기게 된다. 따라서 그룹 최적화 단계에서 이러한 플로우들을 제거한다. 또한 무분별한 시그니쳐 생성을 방지하기 위해 너무 적은 플로우를 포함하는 그룹을 제거한다. 마지막으로 그룹의 대표벡터와 패킷 순서 별 가중치 및 패킷 순서 별 임계치를 요소로 갖는 WPSS 시그니쳐를 추출한다.

오른쪽의 트래픽 분류 단계에서는 입력되는 패킷들을 플로우 형태로 바꾼 뒤 플로우 내 패킷의 전송 순서와 방향, 그리고 페이로드 크기에 시그니쳐의 순서 별 가중치를 곱한 값으로 벡터로 변형하고 시그니쳐에 매칭 시킨다. 매칭 기준은 시그니쳐 생성 단계의 2 차 그룹핑 기준과 같다.

Ⅳ. 실험 및 결과 분석

본 장에서는 제안하는 방법론의 트래픽 분류 결과를 [3]의 분류 결과와 비교하여 성능을 검증한다. 실험은 학내 망 트래픽을 대상으로 하였으며, 시그니쳐 생성과 정확한 검증을 위해서 TMA(Traffic Measurement Agent)[4]를 이용하여 약 10 일간 수집한 정답지 데이터(ground-truth)를 사용하였다.

표 1. 분석률 비교

		flow	packet	byte
[3]	분석률	66.32%	40.75%	37.47%
	정확도	99.68%	98.72%	99.27%
WPSS	분석률	64.09%	47.04%	52.63%
	정확도	99.60%	98.79%	99.46%

표 1 은 [3]의 방법을 적용한 분류 결과와 WPSS 를 적용한 분류 결과로, 분석률과 정확도를 비교한 표이다. 분석률, 정확도 모두 WPSS 가 [3]보다 플로우 단위에서는 조금 낮으며 패킷, 바이트 단위에서는 더 높은 결과를 내었다. WPSS 가 많은 패킷을 포함하는 플로우를 분석하는 성능이 더 뛰어나기 때문이다.

하지만 패킷을 적게 포함한 플로우를 분석하는 성능은 [3]에 비해 조금 뒤쳐진다. 해당 문제는 향후 연구로 남긴다.

Ⅴ. 결론 및 향후 과제

본 논문에서는 플로우 내 패킷의 전송 순서 별 가중치를 이용한 WPSS(Weighted Payload Size Sequence)기반 트래픽 분류 방법론을 제안하였으며 기존의 동일한 Feature 를 가진 [3]과 실험 결과 비교를 통해 그 성능을 검증하였다.

향후 연구에서는 시그니쳐의 충돌 문제를 해결할 수 있는 다양한 특징에 대한 연구를 진행할 계획이다.

참 고 문 헌

[1] SUN Mei-feng, CHEN Jing-tao, “Research of the traffic characteristics for the real time online traffic classification”, The Journal of China Universities of Posts and Telecommunications, June 2011, 18(3): 92-98

[2] 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.

[3] 박진완, 김명섭, "통계 시그니쳐 기반 트래픽 분석 시스템의 성능 향상", KIPSTC,.18C.4., Aug. 2011, pp. 243-250.

[4] 윤성호, 노현구, 김명섭, “ TMA(Traffic Measurement Agent)를 이용한 인터넷 응용 트래픽 분류”, 2008 년 제 29 회 정보처리학회 춘계학술발표대회, 대구, 경일대학교, May, 17, 2008, 제 15 권 제 1 호, pp.946-949..