Flow Similarity 를 활용한 응용 트래픽 분류에 관한 연구

*박병철, *정재윤, **홍원기
*포항공과대학교 컴퓨터공학과, **포항공과대학교 정보전자융합공학부
fates@postech.ac.kr, dejavu94@postech.ac.kr, jwkhong@postech.ac.kr

A Study on the Application Traffic Classification Based on Flow Similarity

*Byungchul Park, *Jae Yoon Chung, **James Won-ki Hong
*Dept. of Computer Science and Engineering, POSTECH
**Division of IT Convergence Engineering, POSTECH

요 약

어플리케이션 트래픽의 정확한 분류는 네트워크의 상태 분석과 관리 측면에서 매우 중요한 부분을 차지하고 있다. 가장 전통적인 트래픽 분류 방법인 well-known 포트 비교 방법은 더 이상 높은 정확도를 보장하지 못하고 있으며, 새로이 소개된 signature 기반의 트래픽 분류 방법, 트래픽 행동 분석 및 기계 학습에 의한 트래픽 분류 방법은 상대적으로 높은 정확도를 보이지만 높은 complexity 를 갖는다. 본 논문에서는 네트워크 flow 간의 similarit 를 활용한 새로운 트래픽 분류 방법을 소개 하고 검증한다.

Ⅰ. 서 론

본 논문에서는 패킷의 페이로드를 벡터로 변환하고, 벡터 시리즈들 간의 유사도 측정결과를 활용하여 트래픽을 분류할 수 있는 새로운 트래픽 분류 방법을 제한다. 해당 트래픽 분류 방법은 자연언어 처리 분야의 문서 분류 [1]에서 널리 활용되고 있는 consine similarity 를 이용한다. 본 논문에서 제시되고 있는 트래픽 분류와 문서 분류는 해당 트래픽과 문서를 벡터로 변환하고 각 벡터들간의 similarity 를 측정하여 이를 기반으로 트래픽과 문서를 분류하는 유사성을 갖고 있지만, 텍스트 문서를 이루는 기본 단위인 word 와 패킷을 이루는 바이너리 데이터의 근본적인 차이 점이 존재한다. 따라서 본 논문에서는 이러한 차이점을 반영하여 패킷 페이로드 데이터를 효과적으로 벡터로 변환하는 방법과 이를 활용한 트래픽 분류 방법을 제시한다.

Ⅱ. 본론

본론에서는 문서 분류 방법의 기술을 적용한 트래픽 분류 방법의 동작과 해당 분류 방법의 정확도 검증에 대하여 기술한다.

트래픽 데이터 간의 Cosine similarity 를 측정하기 위해서는 해당 트래픽을 벡터 형태로 변환하여야 한다. 트래픽-벡터 변환 과정에는 Vector Space Model 이 사용된다. Vector Space Modeling 은 자연 언어 처리 분야에서 텍스트 문서를 벡터로 표현하기 위해 사용하는 algebraic model 로 문서내의 key word 들의 출현 빈도를 측정하여 텍스트 문서를 term-frequency 벡터의 형태로 표한다. 하지만 패킷 데이터상에는 word 라는 형태의 기본 구성 요소가 존재하지 않기 때문에 벡터 변환을 위해 다음과 같이 word 를 정의 한다.

- **정의 1**: *word* 는 i-bytes 의 sliding-window 내에 존재하는 페이로드 데이터를 의미한다.

위의 정의에 따라 패킷 페이로드상의 word 의 출현 빈도를 측정하여 정의 2 와 같이 term-frequency 벡터의 형태로 패킷을 표현할 수 있다.

- **정의 2**: $Payload\ Vector = [w_1 w_2 \ldots w_n]^T$ (w_i 는 i 번째 word 의 출현 횟수를 의미하며 n 은 전체 word 의 수를 나타낸다.)

본 연구에서는 페이로드 데이터의 벡터 형태로 표현할 수 있는 가장 단순한 형태로 window 의 사이즈를 2 로 설정하였다. 만약 window 사이즈가 커진다면 벡터의 차수가 기하급수적으로 증가하며, 너무 작을 때에는 페이로드 데이터의 순차적 형태를 표현하기가 어렵다. Window 사이즈가 2 일 때 페이로드 벡터의 차수는 전체 word 의 수 (2^{8*2})와 같은 65536 를 갖는다.

패킷 페이로드 데이터가 벡터로 변환이 되면 벡터들간의 similarity 를 계산 함으로써 패킷들간의 similaitry 를 수치적으로 계산할 수 있다. 본 연구에서는 다양한 similarity metric 들 중 하나인 cosine similarity 를 패킷들 간의 similarity 측정에 사용하였다. Cosine similarity 는 두 벡터가 이루는 각도의 cosine

본 연구는 지식경제부 및 한국산업기술평가관리원의 산업원천기술개발사업(정보통신)의 일환으로 수행하였음.
[KI0035594 , HiMang(Highly Manageable Network and Service Architecture for New Generation) 원천 연구]

값을 계산 하며, 그 값은 -1 (두 벡터가 반대일 경우)부터 1 (두 벡터가 동일할 경우)의 범위 내에 분포한다. Payload vector 의 정규화를 통해 similarity 값을 단순히 두 벡터의 내적 값을 계산할 수 있으며, 두 패킷이 유사할수록 similarity 값은 1 에 가까운 값을 갖는다. 수식 (1)은 cosine similarity 의 수학적인 정의이다.

$$Similarity(p_1, p_2) = \frac{V(p_1) \cdot V(p_2)}{|V(p_1)||V(p_2)|} \quad (1)$$

Flow 는 동일한 IP 주소, 포트, 프로토콜을 공유하는 패킷들의 집합이므로 패킷을 벡터의 형태료 표현 가능하다면 flow 를 matrix 형태로 정의하는 것이 가능하다. Payload flow matrix (PFM) 다음과 같이 정의된다.

- **정의 3**: $PFM = [\overrightarrow{p_1}\ \overrightarrow{p_2} \dots \overrightarrow{p_k}]^T$
  ($\overrightarrow{p_i}$는 정의 2 의 payload vector 를 의미한다)

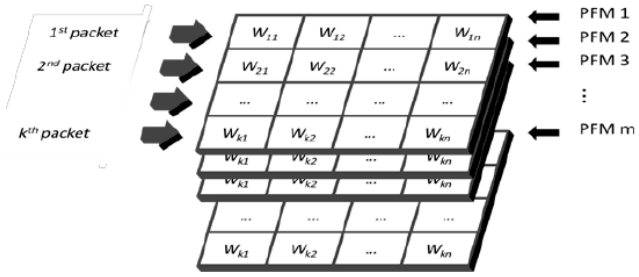

그림 1. Collectec PFM

그림 1 은 collect PFM 을 나타내고 있다. 그림의 각 레이어는 어플리케이션 트래픽의 flow 를 의미한다. 위와 같이 수집된 PFM 은 새로운 어플리케이션 signature[3]로 활용되며, 분류할 트래픽이 트래픽 분류 시스템이 진입되었을 때 수집된 PFM 과의 similarity 값의 계산을 통해 해당 어플리케이션으로 트래픽이 분류된다. 그림 2 는 flow 단위의 similarity 측정 과정을 나타낸다.

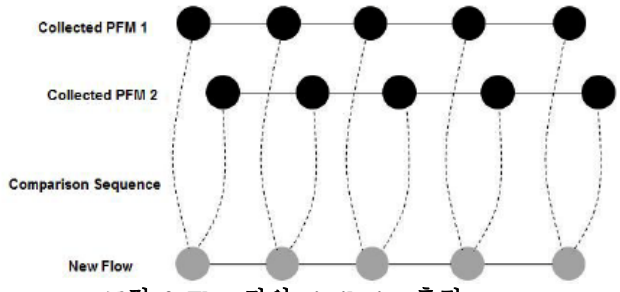

그림 2. Flow 간의 similarity 측정

Flow 상의 패킷들은 각각 상응하는 collected PFM 의 payload vector 와의 cosine similarity 를 계산하게 되며 flow 의 similarity 는 각 payload vector 들의 similarity 의 합으로 측정이 된며, 측정된 similarity 값에 의해 분류된 flow 의 어플리케이션 이름이 결정되고 트래픽이 분류된다.

제시된 트래픽 분류 방법의 정확도를 검증하기 위해 실제 캠퍼스 네트워크에서 발생된 트래픽에 flow similarity 를 통한 트래픽 분류 방법을 적용하였다. 목표 어플리케이션은 국내외에서 많이 사용되고 P2P 어플리케이션 3 종 (BitTorrent[4], LimeWire, FileGuri)과 YouTube[5]가 선정되었다.

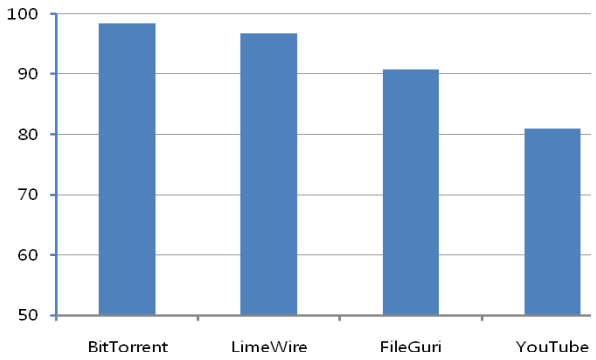

그림 3. 트래픽 분류 정확도 (%)

그림 3 은 각 어플리케이션들의 분류 정확도를 나타내고있다. YouTube 의 트래픽은 일반적인 HTTP 트래픽과 거의 동일한 패킷 데이터를 갖기 때문에 정확도가 다른 어플리케이션에 비해 낮았지만 이를 제외한 어플리케이션의 정확도는 모두 90%이상으로 signature 를 활용한 트래픽 분류 정확도[2]와 거의 유사한 정확도를 보인다.

Ⅲ. 결론

본논문에서는 패킷 페이로드 데이터간의 similarity 를 활용한 새로운 트래픽 분류 방법을 제시하고 분류 정확도를 실제 트래픽을 통해 검증하였다. 전체 트래픽 분류 정확도는 96%를 보였으며 이 결과는 현재 가장 정확한 트래픽 분류 방법으로 알려진 signature 를 활용한 트래픽 분류의 정확도와 거의 유사한 정확도이다.

향후 연구로 분류 가능한 어플리케이션의 수를 증가시키고 실시간 분류 시스템 개발을 계획하고 있으며, 제시된 트래픽 분류 방법을 암호화된 트래픽의 분류로 확장하려고 한다.

참 고 문 헌

[1] G. Salton and C. Buckley, "Term-weighting Approaches in Automatic Text Retrieval", Information Processing and Management, 1988, vol. 24, No. 5, pp. 513-523.

[2] Byungchul Park, Young J. Won, Myung-Sup Kim, and James W. Hong. "Towards Automated Application Signature Generation for Traffic dentification", IEEE/IFIP Network Operations and Management Symposium (NOMS 2008), Salvador, Brazil, Apr. 7-11, 2008, pp. 160-167.

[3] T. Karagiannis, K. Papagiannaki, and M. Faloutsos, "BLINC: Multilevel Traffic Classification in the Dark", ACM SIGCOMM 2005, Philadelphia, PA, USA, Aug. 21-26, 2005.

[4] BitTorrent, http://www.bittorrent.com/.

[5] YouTube, http://youtube.com/.