네트워크 플로우의 연관성 모델을 이용한 트래픽 분류방법

구영훈 ○ 이성호 정우석 김성민 김명섭

고려대학교

{ gyh0808, gaek5, hary5832, gogumiking, tmskim }@korea.ac.kr

A Method for Traffic Classification Using Correlation of Network Flow

Young-Hoon Goo ○ Sung-Ho Lee Woosuk Jung Sung-Min Kim Myung-Sup Kim

Korea Univ.

요 약

오늘날의 네트워크는 고속화와 유비쿼터스 환경으로 인해 다양한 응용이 급속도로 생성되고 있으며 네트워크 트래픽도 매우 복잡해지고 있다. 이에 효율적인 네트워크 운용 및 관리를 위한 구체적인 단위의 트래픽 분류가 필수적이다. 다양한 트래픽 분류 방법이 연구되고 있는 가운데 아직 트래픽을 완벽하게 분류해내는 방법론은 개발되지 않은 실정이다. 이에 본 논문에서는 네트워크 플로우의 연관성 모델을 정의하고 이를 기반으로 트래픽을 분류하는 방법을 제안하고자 한다. 그리고 실험을 통해 본 분류 방법론이 높은 정확도와 분석률의 방법론이라는 것을 증명한다.

1. 서 론

오늘날의 네트워크는 고속화와 유비쿼터스 환경으로 인하여 다양한 응용이 급속도로 생성되고 있으며 이에 따라 네트워크 트래픽도 매우 복잡해지고 있다. 이러한 상황 속에서 네트워크 트래픽 모니터링과 분석은 효율적인 네트워크 운영과 안정적 서비스 제공에 있어 필수적이다. 인터넷 트래픽의 정확한 응용별 분류는 다양한 트래픽 분석 요구에 대처하기 위해 반드시 선행되어야만 한다.

다양한 트래픽 분류 방법이 연구되고 있는 가운데 아직 트래픽을 완벽하게 분류해내는 방법론은 개발되지 않은 실정이다. 먼저, 헤더 시그니쳐 기반 분류 방법[1]은 다양한 프로토콜을 사용하는 응용, 둘 이상의 포트 또는 임의의 포트를 설정할 수 있는 기능을 제공하는 응용 등 현대의 복잡한 구조를 갖는 응용에 대해서는 신뢰성을 갖기 힘들다. 페이로드 시그니쳐 기반 분류 방법은 분석률과 정확도 측면에서 가장 높은 분석 성능을 보이지만 시그니쳐 추출 작업이 수작업으로 이루어져 응용 프로그램의 변화에 신속하게 대처할 수 없으며, 암호화된 패킷에 대해서는 트래픽을 분류할 수 없다. 이를 극복하기 위한 통계 시그니쳐 기반 분류 방법[2]은 생성시 전문성을 가진 인력과 많은 시간이 필요하며 시그니쳐가 생성자의 능력에 큰 차이가 있을 수 있다. 또한 암호화된 트래픽은 분류할 수 없는 큰 단점을 가지고 있다. 통계 기반 시그니쳐 분류 방법은 암호화된 트래픽을 분류할 수 있지만, 통계 정보를 사용할 경우, 특정 응용 프로그램에서 사용한 엔진 또는 응용 프로토콜에 의존적인 시그니쳐가 생성될 가능성이 높다. 기계학습 기반 트래픽 분류 방법은 많은 양의 샘플 데이터 확보가 정확도에 큰 영향을 미치며 동일한 응용 계층 프로토콜을 이용하는 응용에 대해서는 분류가 어려운 문제가 발생된다.

본 논문에서는 네트워크 플로우의 연관성 모델을 정의하고 이를 기반으로 트래픽을 분류하는 방법을 제안한다. 제안하는 방법론은 입력 트래픽에 대해 연관성을 자동으로 계산하여 그룹화하므로 분류 속도가 빠르고 암호화된 트래픽 분석이 가능하며 연쇄적인 그룹화로 분석률을 최대화할 수 있다.

본 논문은 본 장의 서론에 이어 2장에서 제안하는 방법론에 대해 기술하고, 3장에서는 본 방법론의 효과적인 적용 방안에 대해 기술한다. 4장에서는 본 방법론의 타당성을 실험을 통해 증명하고, 마지막으로 5장에서는 결론 및 향후 연구를 기술한다.

2. 네트워크 플로우의 연관성 모델을 이용한 트래픽 분류 방법

본 장에서는 제안하는 방법론을 기술한다. 네트워크 플로우의 연관성 모델은 크게 유사성 모델과 연관성 모델로 구성된다. 그림 1은 제안하는 방법론의 분류 체계도이다. 먼저 타겟 트래픽과 타겟 트래픽에서 분류하고자 하는 탐지 대상(응용 또는 악성행위)의 플로우를 Seed Group으로 선정하여 유사성 모델에 입력한다. Seed Group은 악성행위의 경우 IDS 또는 IPS의 경보, 응용의 경우 다양한 시그니쳐를 통해 얻은 정보를 통해 선정할 수 있다. 유사성 모델에서는 Seed Group과 타겟 트래픽과의 유사성을 계산하여 일정 임계치 이상을 그룹화하여 Group1을 출력한다. 다음으로

* 이 논문은 2015년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No.2015R1D1A3A01018057).

Group1을 연결성 모델의 입력으로 사용하면 Group1과 그룹화되지 않은 트래픽과의 연결성을 계산하여 Group2를 출력한다. 더 이상 그룹화되지 않을 때까지 해당 출력을 반복적으로 연결성 모델에 입력한다. 최종적으로 Seed Group과 Group1부터 GroupN까지의 트래픽을 분류 결과로 출력한다.

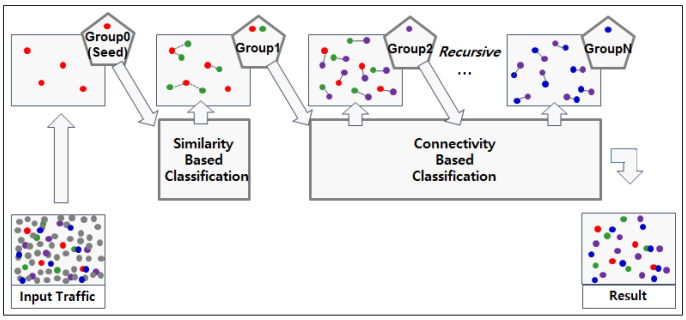

그림 1. 트래픽의 연관성을 이용한 트래픽 분류 체계도

플로우간 유사성은 플로우의 통계정보들을 벡터로 구성하여 유클리드 거리값으로 계산한다. 사용되는 통계정보들은 플로우를 구성하는 모든 패킷의 inter-arrival time, 처음 16개 패킷간 inter-arrival time, 모든 패킷 페이로드 크기, 첫 5개 패킷의 페이로드 크기의 최대값, 최소값, 평균, 표준편차 총 16개의 속성들과 첫 5개 패킷의 방향성을 고려한 페이로드 크기를 사용하여 총 21가지의 속성을 사용한다. 수식 1은 플로우 f_x와 f_y의 유사성을 계산하는 식이다. a_i는 각 속성을 의미하며 각 속성들의 가중치를 동일하게 하기 위하여 Min-Max 정규화를 각 속성에 적용하여 0~1 사이의 값으로 척도를 맞춘다. 계산되는 유클리드 거리를 다시 Min-Max 정규화를 통해 0~1 사이의 값으로 나타내고 비유사도를 유사도로 전환하기 위해 1에서 해당 값을 뺀다. 최종적으로 플로우간 유사성은 0~1 사이의 값으로 표현되고 유사할수록 1에 가까운 값을 가진다. 유사성 모델에서는 Seed Group과 타겟 트래픽의 유사성을 계산하여 특정 임계치를 이상의 유사성을 가진 플로우들을 그룹화한다.

$$Sim(f_x, f_y) = 1 - \frac{Euclidean\ distance}{\sqrt{21}} = 1 - \frac{\sqrt{\sum_{i=1}^{21}(a_i(f_x) - a_i(f_y))^2}}{\sqrt{21}}$$

$$, where\ 0 \le a_i(f_x, f_y) \le 1$$

수식 1. 플로우간 유사성 계산

플로우간 연결성은 플로우의 발생시간, 출발지 IP 주소와 도착지 IP 주소쌍, 출발지 포트와 도착지 포트쌍, L4 프로토콜 총 4개의 속성 값을 이용하여 계산한다. 수식 2는 플로우 f_x와 f_y의 연결성을 계산하는 식이다.

$$Con(f_x, f_y) = \sum_{i=1}^{4}(w_i \times a_i(f_x, f_y)),\ where\ 0 \le a_i(f_x, f_y) \le 1, \sum_{i=1}^{4} w_i = 1$$

수식 2. 플로우간 연결성 계산

w_i는 속성별 가중치를 나타낸다. 가중치의 합은 1이며 탐지 대상의 종류에 따라 발생하는 트래픽이 상이할 수 있음을 염두하여 속성들의 가중치를 조정할 수 있다. a_i는 연결성의 각 속성을 의미한다.

$$a_{ST}(f_x, f_y) = 1 - \sqrt{\frac{dist(f_x, f_y)}{maxdist(F)}}$$

수식 3. 시작시간 연결성 계산

수식 3은 플로우간 시작시간의 연결성을 계산하는 식이며 min-max 정규화를 통해 0~1 사이의 값이 나오게 된다. $dist(f_x, f_y)$는 플로우의 시작시간 차를 의미하며 $maxdist(F)$는 타겟 트래픽의 전체 플로우들의 시작시간 차 중 최대값을 의미한다.

$$a_{IP}(f_x, f_y) = \left(\frac{prefixlen_{src}(f_x, f_y)}{32}\right)^2 + \left(\frac{prefixlen_{dst}(f_x, f_y)}{32}\right)^2$$

수식 4. 출발지, 도착지 IP 주소쌍 연결성 계산

$$a_{PT}(f_x, f_y) = \left(\frac{prefixlen_{src}(f_x, f_y)}{16}\right)^2 + \left(\frac{prefixlen_{dst}(f_x, f_y)}{16}\right)^2$$

수식 5. 출발지, 도착지 포트쌍 연결성 계산

수식 4와 5는 각각 출발지, 도착지 IP주소쌍 연결성과 출발지, 도착지 포트쌍 연결성을 계산하는 수식이다. 수식4의 $prefixlen_{src}(f_x, f_y)$는 f_x의 출발지 주소와 f_y의 출발지 주소 32비트 중 앞부분이 같은 비트의 개수와 f_x의 출발지 주소와 f_y의 도착지 주소 중 큰 값을 의미하며, $prefixlen_{dst}(f_x, f_y)$는 f_x의 도착지 주소와 f_y의 도착지 주소 32비트 중 앞부분이 같은 비트의 개수와 f_x의 도착지 주소와 f_y의 출발지 주소 중 큰 값을 의미한다. 수식 5도 수식4의 계산식과 마찬가지로 포트쌍의 16비트의 앞부분이 얼마나 같은지를 계산하는 식이며, 이는 한 응용이 실행되었을 때 발생하는 통신 플로우의 포트는 서로 유사한 범위의 포트를 사용한다는 점에 착안하였다.

$$a_{PROT}(f_x, f_y) = \begin{cases} 0: f_x.PROT \neq f_y.PROT \\ 1: f_x.PROT = f_y.PROT \end{cases}$$

수식 6. 프로토콜 연결성 계산

수식 6은 프로토콜 연결성을 계산하는 식으로 프로토콜이 같으면 1 아니면 0으로 계산된다. 최종적으로 플로우간 연결성은 0~1 사이의 값으로 표현되고 유사할수록 1에 가까운 값을 가진다. 연결성 모델에서는 유사성 모델에서 그룹화된 Group1과 타겟 트래픽의 연결성을 계산하여 특정 임계치 이상의 연결성을 가진 플로우들을 그룹화하고, 이 그룹을 다시 연결성 모델에 반복적으로 입력하여 탐지 대상의 트래픽들을 계속하여 그룹화할 수 있다.

3. 네트워크 플로우의 연관성 모델을 이용한 트래픽 분류 방법의 효과적인 적용 방안

유사성 모델은 통계 정보를 이용하고 연결성 모델은 헤더정보 및 플로우 발생 시간을 이용한다는 점에서 통계 시그니쳐 기반 분류 방법과 헤더 시그니쳐 기반 분류 방법과 비슷하다. 하지만, 제안하는 방법론은 유사성 모델을 먼저 1차적으로 사용하고 이에 분류하지 못한 나머지 트래픽을 연결성 모델을 연쇄적으로 사용함으로써 분석률을 최대화할 수 있으며 암호화된 트래픽 또한 분류가 가능하다. 또한

연결성 모델에서는 헤더 시그니쳐처럼 특정 IP 리스트와 포트, 프로토콜이 같은 플로우만 분류하는 것이 아닌 연결성의 계산을 통하여 기준 플로우의 헤더정보와 유사한 플로우를 그룹화하며 1차적으로 유사성 모델의 출력을 입력으로 사용하기 때문에 포트 기반 분류 방법의 한계를 극복할 수 있다.

본 방법론을 효과적으로 적용하기 위해서는 탐지 대상별로 유사성 모델의 가이드라인과 연결성 모델의 가이드라인을 설정해 놓고 이에 따라 분류하여야 한다. 유사성 모델의 탐지 대상별 가이드라인이란 어떤 플로우를 Seed Group으로 사용해도 정확도가 100%가 되는 정교한 임계치를 의미하며, 연결성 모델의 탐지 대상별 가이드라인이란 분석률을 최대화할 수 있도록 탐지 대상의 특성을 잘 반영한 연결성의 속성별 가중치와 임계치를 의미한다. 이와 같이 가이드라인이 필요한 이유는 다음과 같다.

먼저, 유사성 모델에서는 그룹화된 플로우들을 연결성 모델의 입력으로 사용하며 이를 연쇄적으로 사용하기 때문에 유사성 모델에서 False Positive가 조금이라도 있을 경우 연결성 모델에서는 반복되는 횟수에 따라 False Positive가 급증하게 된다. 따라서, 유사성 모델에서는 반드시 정확도가 100%가 되어야 한다. 때문에, 탐지 대상별로 어떤 Seed를 사용하여도 False Positive가 0이 나오는 정교한 임계치가 설정되어야 한다. 두번째로 연결성 모델에서는 응용 또는 공격의 종류에 따라 발생하는 트래픽의 특성이 상이할 수 있음을 염두하여 가중치를 조정할 수 있게함으로써 보다 더 정확도와 분석률을 향상시킬 수 있다.

4. 실험 및 결과

본 장에서는 본 방법론이 높은 정확도와 분석률을 갖는 방법론임을 실험을 통해 증명한다.

실험 환경은 다음과 같다. 먼저 여러 호스트에서 탐지하고자 하는 대상의 정답지 트래픽과 노이즈 트래픽을 수집한다. 본 실험에서는 4개의 호스트에서 Torrent 트래픽과 이에 대한 노이즈 트래픽을 수집하였다. 다음으로 정답지 플로우 중 하나를 Seed로 선택하고 유사성 모델의 임계치를 0~1까지 0.0001씩 증가시키며 분류를 한다. 결과에서 False Positive가 0이 되는 임계치를 확인한다. 모든 정답지 플로우를 Seed로 선정하여 위의 과정을 반복한 후 모든 Seed에 대하여 False Positive가 0이 되는 임계치의 최대값을 확인한다. 임계치가 높을수록 분석률은 감소하지만 정확도는 향상되기 때문에 이 임계치는 최대값이 된다. 이 유사성 모델 임계치의 최대값을 사용하여 유사성 모델로 1차적으로 분류를 한 후 연결성 모델에 입력으로 넣고 연결성 모델 속성의 가중치와 연결성 모델의 임계치를 0.1씩 조정한 조합의 모든 경우의 수로 그룹화된 결과를 확인하고 분석률이 최대이며 정확도는 100%인 조합을 확인한다. 본 실험을 각각 다른 Torrent 트레이스를 사용하여 4회 반복함으로써 Torrent의 강건한 유사성 및 연결성 가이드라인이 존재함을 확인한다. 표 1은 실험에 사용된 트래픽의 정보이며 표2는 실험 결과다.

Experiment#		Flow	Packet	Byte
1	Torrent	582	237205	222.6MB
	Noise	3189	85747	138.3MB
2	Torrent	924	173018	161.9MB
	Noise	2573	197871	336.7MB
3	Torrent	732	182642	163.6MB
	Noise	2669	257903	486.9MB
4	Torrent	579	230648	208.4MB
	Noise	2486	285159	525.5MB

표 1. 실험에 사용된 트래픽의 정보

Ex	Sim_t	Con_t	w_{ip}	w_{po}	w_{pr}	w_t	TP	FP
1	0.996	0.6	0.2	0.4	0.1	0.3	93.04%	0%
2	0.9759	0.6	0.2	0.4	0.1	0.3	99.20%	0%
3	0.9949	0.6	0.2	0.4	0.1	0.3	99.78%	0%
4	0.9870	0.6	0.2	0.4	0.1	0.3	95.36%	0%

표 2. 실험 결과

표 2에서 1열은 실험번호. 2열~7열은 유사성 모델 임계치, 연결성 모델 임계치와 속성별 가중치를 나타내며 8열은 True Positive, 9열은 False Positive를 의미한다. 4번의 실험 결과 Torrent의 유사성 모델의 정확도 100%의 임계치는 0.996임을 확인되었다. 또한 연결성 모델을 1회만 사용하였음에도 불구하고 높은 분석률을 보였으며 반복적으로 사용할 경우 더욱 높을 것으로 예상된다. 4번의 실험 결과 정확도는 100%이며 분석률을 최대로 하는 연결성 모델 임계치와 속성별 가중치 조합이 모두 0.6, 0.2, 0.4, 0.1, 0.3로 일정한 것으로 보아 Torrent의 가이드라인을 확인하였으며, 다른 탐지 대상에 대해서도 강건한 가이드라인이 존재할 것으로 예상된다.

5. 결론 및 향후 연구

본 논문에서는 네트워크 플로우의 연관성 모델을 이용한 트래픽 분류 방법을 제안하고 본 방법론의 타당성을 실험을 통해 증명하였다. 향후 연구로는 대표 응용을 선정하여 가이드라인을 찾고 본 방법론을 악성 탐지에 적용하는 연구를 진행할 계획이다.

참고문헌

[1] S.-H. Yoon, and M.-S. Kim, “Research on Header Signature Maintenance Method for Internet Application Traffic Identification.” In Proc KICS ICC 2011, pp.1200-1201, Jeju Island, Korea, June 2011

[2] H.-M. An, M.-S. Kim. “Performance Improvement of the Statistical Information Based Traffic Identification System”, KTCCS, vol. 2, no. 8, pp.335-342, Aug. 2013.