서비스 별 트래픽 분류를 위한 클러스터링 기반 페이로드 시그니쳐 자동 생성 연구

심규석, 구영훈, 박지태, 김명섭
고려대학교
{kusuk007, gyh0808, pjj5846, tmskim}@korea.ac.kr

A Study on the Automatic Payload Signature Generation based on Clustering for Service-Specific Traffic Classification

Kyu-Seok Shim, Young-Hoon Goo, Jee Tae Park, Myung-Sup Kim
Korea Univ.

요 약

오늘날 네트워크 자원을 활용하는 응용이 증가되고, 사용자의 요구가 증가되면서 응용에서 사용할 수 있는 서비스도 증가되고 있다. 효율적인 네트워크 관리를 위해서는 응용에서 사용되는 서비스 별로 분류할 수 있는 방안이 요구된다. 그러나 현재 시그니쳐 자동 생성에 관한 연구는 응용을 분류하는 시그니쳐 추출 단계에서 정체되어 있고, 응용에서 사용되는 서비스 별로 분류할 수 있는 시그니쳐를 자동으로 추출하기 어렵다. 현재까지 서비스 별로 분류하기 위한 시그니쳐를 추출하는 방법은 전처리 단계에서 서비스 별 트래픽을 수집하거나, 후처리단계에서 서비스 별 시그니쳐를 분류하는 방법으로 나누어지고 있다. 본 논문에서는 전, 후처리 단계없이 응용 트래픽 수집 후 해당 트래픽을 군집화하여 서비스별로 나누고 군집화 된 트래픽으로부터 시그니쳐를 추출하는 방법을 제안한다. 본 방법론으로 인해 서비스 별로 시그니쳐를 추출할 수 있는 기대효과를 나타낼 수 있다.

Ⅰ. 서 론

오늘날 대부분의 응용은 네트워크 자원을 활용하고, 네트워크 자원의 활용도가 증대되면서 네트워크 관리 대상 트래픽양은 기하급수적으로 증대되고 있다. 뿐만 아니라 이러한 응용들은 사용자의 요구에 따라 응용에서 사용되는 서비스들의 종류를 증가시키고 있다. 네트워크 관리자 입장에서는 응용 별로 트래픽을 분류할 수 있지만, 응용에서 발생하는 서비스 별로 트래픽을 분류할 수 없다. 따라서 효율적인 네트워크 관리가 어려워지고 있다.

현재까지 서비스 별로 분류하기 위한 시그니쳐를 추출하는 방법은 전처리 단계에서 서비스 별 트래픽을 수집하거나, 후처리단계에서 서비스 별 시그니쳐를 분류하는 방법으로 나누어지고 있다. 이러한 방법은 네트워크 관리자가 직접 전처리과정에서 분류하거나, 후처리과정에서 추출된 시그니쳐를 봐야만 진행할 수 있는 과정이다. 그러나, 네트워크 자원을 활용하는 응용과 응용에서 사용되는 서비스들의 개수는 셀 수 없을 정도로 많이 존재하고, 새롭게 생성되고 있다. 따라서 기존 방법으로 서비스별 트래픽 분류는 매우 어렵다.

따라서 본 논문에서는 트래픽 군집화를 통한 서비스 별 페이로드 시그니쳐 자동 생성 방법을 제안한다. 응용 트래픽을 수집하고, 수집된 트래픽을 군집화를 통해 서비스 별로 구분한 뒤, 군집 집합 별로 시그니쳐를 생성함으로써 서비스 별 페이로드 시그니쳐를 생성할 수 있다. 군집 집합 별 페이로드 시그니쳐를 생성하는 과정에서 페이로드 시그니쳐 자동 생성 방법을 적용하여 보다 손 쉽고 정확한 시그니쳐를 생성한다.

본 논문은 본장 서론에 이어, 2장 본론에서 트래픽 군집화를 통한 서비스 별 페이로드 시그니쳐 자동 생성 방법을 제안한다. 마지막으로 3장에서 결론 및 향후연구에 대해 언급하고 본 논문을 마친다.

본 논문은 2015년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(No.2015R1D1A3A01018057)과 2016년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행 된 과학기술인문융합연구사업임(No.NRF-2016M3C1B6929228).

Ⅱ. 본 론

서비스 별 페이로드 시그니쳐 자동 생성 시스템은 다음 그림1과 같이 구성된다. 먼저, 정답지 트래픽 수집부에서는 응용의 정답지 트래픽을 TMA를 이용해서 수집하고, 트래픽 군집부에서는 클러스터링 기법을 이용하여 각 응용 트래픽 트레이스를 서비스별로 구분한다. 구분된 트래픽은 페이로드 시그니쳐 자동 생성 시스템으로 입력되어, 서비스 별 시그니쳐 추출이 가능해진다.

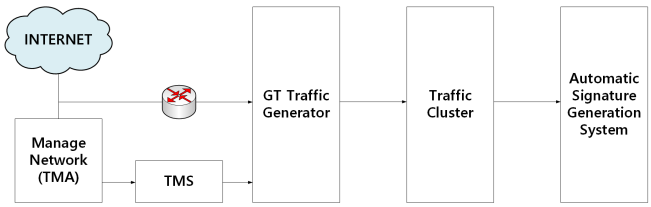

그림 1. 서비스 별 페이로드 시그니쳐 자동 생성 시스템

정답지 트래픽 수집부(GT Traffic Generator)에서는 TMA(Traffic Measurement Agent)와 TMS(Traffic Measurement Server)를 통해 수집될 수 있는 로그 데이터와 네트워크 미러링을 통해 수집된 트래픽 데이터의 매칭을 통해 정답지 트래픽을 수집한다. 그림2와 같이 로그데이터에는 트래픽 사용의 출처인 프로세스 이름과 트래픽 데이터와 비교할 수 있는 5-tuple(출발지 IP 주소/포트번호, 목적지 IP 주소/포트번호, 프로토콜)에 대한 정보가 포함되어 있다. 따라서 두 개의 데이터 매칭을 통해 프로세스 정보가 포함된 네트워크 트래픽 데이터를 수집할 수 있다.

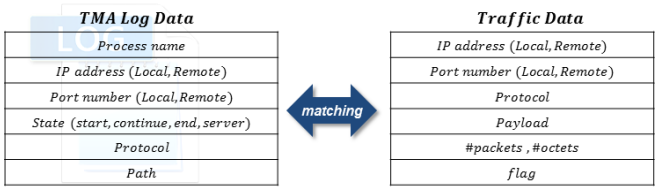

그림 2. TMA 로그데이터와 트래픽 데이터 매칭을 통한 정답지 트래픽 추출

트래픽 군집부(Traffic Cluster)에서는 정답지 트래픽 수집부에서 수집된 정답지 응용 트래픽을 각 플로우 별로 군집화하여 패킷의 통계정보를 기반으로 군집화한다. 서비스 별 페이로드 시그니쳐 자동 생성의 핵심 부분은 수집된 응용 트래픽에서 군집화를 통해 서비스 별 트래픽으로 분류하는 연구이다. 그림 3과 같이 페이로드 시그니쳐 자동 생성 시스템으로 군집화된 트래픽이 분류되서 입력된다.

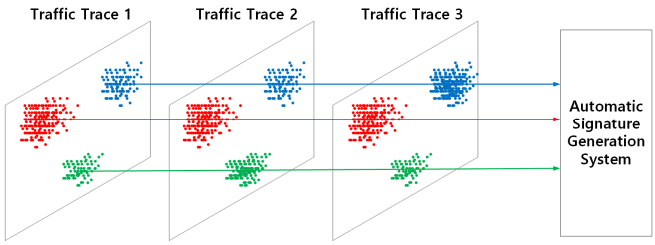

그림 3. 트래픽 군집화를 통한 시그니쳐 자동 생성

기존 방법은 트래픽 군집화 없이 모든 Traffic Trace에서 공통으로 발생한 시그니쳐가 추출된다. 따라서 부정확한 시그니쳐가 추출되고, 추출된 시그니쳐는 해당 응용에 대한 시그니쳐로 분류된다. 하지만 응용에서 사용되는 서비스 별 시그니쳐 추출하기 위해 본 과정이 필수적으로 이루어져야 하며, 본 과정이 이루어짐으로써 부정확한 시그니쳐가 추출되는 경우가 감소될 수 있다.

트래픽 군집화를 위해 트래픽의 Feature를 추출하는 과정이 요구된다. 따라서 트래픽 군집부에서는 트래픽 Feature 추출 과정이 포함된다. 본 연구에서 사용되는 트래픽 Feature는 플로우의 크기, 플로우에 포함된 패킷의 개수, 플로우 발생 시간(Duration), 지연시간 등이 사용된다. 다음의 Feature들을 기준으로 비슷하게 발생하는 플로우들을 군집화한다. 예를 들면, 동영상 서비스의 경우 플로우의 크기가 큰 플로우들만이 모여져서 동영상 서비스에 대한 시그니쳐가 추출될 수 있다.

그러나, 시그니쳐 자동 생성에서 추출될 결과물에 해당 시그니쳐는 어떤 응용의 어떤 서비스를 명시해줘야한다. 응용은 정답지 트래픽을 통해 해결되지만 서비스는 해당 과정에서 명시할 수 없다. 본 연구에서 이러한 응용 및 서비스 명시는 라벨이라고 정의한다. 라벨은 추출된 시그니쳐로 트래픽을 분류 시 해당 트래픽은 “A응용의 B서비스” 라는 것을 명시해주는 필드이다. 서비스의 라벨을 정의하기 위해 본 연구에서는 DNS(Domain Name Server) 응답패킷을 사용한다. 응답패킷의 정보는 다차원 주소 체계를 가지기 때문에 해당 정보에서 서비스 키워드를 추출할 수 있다.

시그니쳐 자동 생성부(Automatic Signature Generation System)에서는 군집화된 트래픽을 입력받아 Apriori 알고리즘을 이용하여 시그니쳐를 자동으로 추출한다[1,2]. 해당 부에서는 총 3단계의 페이로드 시그니쳐를 추출하는데 첫 번째로 콘텐츠 시그니쳐는 트래픽 페이로드를 기준으로 그림2에서 각 Traffic Trace에서 공통으로 발생하는 연속된 문자열을 의미한다. 따라서 하나의 패킷 또는 플로우에서 여러 개의 콘텐츠 시그니쳐가 추출될 수 있다. 두 번째, 패킷 시그니쳐는 동일한 패킷에서 발생하는 콘텐츠 시그니쳐의 집합을 의미한다. 패킷 단위의 정교한 시그니쳐를 생성함으로써 시그니쳐의 정확도를 향상시키고, 패킷 단위에서 중단하는 것이 아닌 세 번째, 플로우 시그니쳐를 생성한다. 플로우 시그니쳐는 동일한 플로우에서 발생하는 패킷 시그니쳐의 집합을 의미한다.

위의 과정을 모두 데이터 마이닝의 한 종류인 Apriori 알고리즘을 사용함으로써 전처리 및 후처리 과정을 생략하고, 모든 문자열을 한번에 비교할 수 있는 메커니즘을 개발한다. 그림4는 패킷 페이로드에서 Apriori 알고리즘을 사용하여 각 타입 별 시그니쳐를 추출하는 과정을 나타낸다.

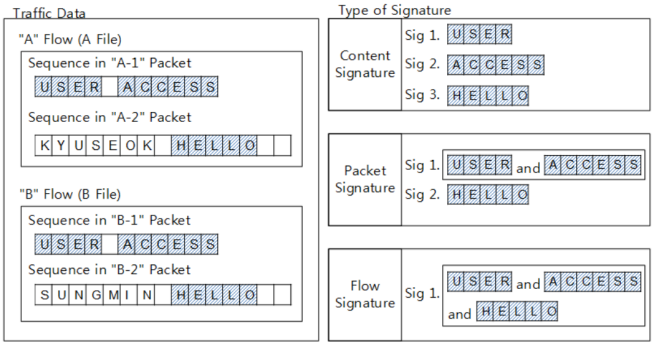

그림 4. 타입별 시그니쳐 생성

Ⅲ. 결론 및 향후연구

본 논문에서는 네트워크 트래픽 서비스 별 분류를 위한 페이로드 시그니쳐 자동 생성 방법을 제안 한다. 본 방법론은 정답지 트래픽 생성부, 트래픽 군집화부, 시그니쳐 자동생성부로 나누어져있다. 정답지 트래픽 생성부에서는 TMA로그데이터를 이용하여 프로세스별로 트래픽을 분류한다. 트래픽 군집화부에서는 클러스터링 기법을 통해 각 트래픽 트레이스에서 플로우별로 트래픽 통계정보를 이용하여 군집화한다. 군집화된 플로우는 각 서비스별 플로우로 구분될 수 있으며, 서비스별 플로우를 이용하여 자동 생성 시스템에 입력된다. 입력된 플로우는 Apriori 알고리즘을 이용하여 콘텐츠, 패킷, 플로우 시그니쳐로 추출되며, 해당 시그니쳐들은 각 응용의 서비스 별로 트래픽 분류가 가능한 시그니쳐들이다.

향후 연구로는 해당 방법론에 대해 각 응용 별로 수집하여 실험을 진행하고, 최적의 클러스터링 기법에 대해 연구를 진행한다. 또한, 클러스터링에서 사용되는 트래픽의 Feature 종류를 선정하여 최적의 결과를 도출하는 연구를 진행할 계획이다.

참 고 문 헌

[1] 심규석, 구영훈, 이성호, Baraka D. Sija, 김명섭, "최신네트워크응용분류를위한자동화페이로드시그니쳐업데이트시스템",통신학회논문지Vol.42 No.01, Jan. 2017, pp. 1-10

[2] 심규석, 김종현, 김성민, 김명섭, "급변하는네트워크트래픽을효율적으로분류하기위한응용시그니쳐자동생성시스템에대한연구",2016년도정보과학회동계종합학술발표회, 휘닉스파크, 강원, Dec. 21-23, 2016, pp.1021-1023

[3] 이수강, 윤성호, 심규석, 김명섭, "인터넷 서비스 식별을 위한 헤더정보 기반 자동 시그니쳐 명명 시스템", 2015년도 한국통신학회 하계종합학술발표회, 라마다호텔, 제주도, Jun. 23-25, 2015.