인터넷 트래픽 분석을 위한 자동화된 시그니쳐 추출 시스템 설계

윤성호, 박준상, 함재현, 김명섭
고려대학교

{sungho_yoon, jungsang_park, jaehyun_ham, tmskim}@korea.ac.kr

Design of Automatic Signature Extraction System for Internet Traffic Identification

Sung-Ho Yoon, Jun-Sang Park, Jaehyun Ham, Myung-Sup Kim
Korea Univ.

요 약

네트워크 기술의 발전과 인터넷 사용의 대중화로 다양한 응용이 출현하고 있다. 효과적인 망 관리를 위해 시그니쳐 기반 트래픽 분석 방법론이 제안되고 있지만, 많은 부분을 수작업에 의존하고 있기 때문에 시시각각으로 생성되고 소멸되는 응용 트래픽을 분석하기에는 많은 어려움이 존재한다. 따라서 본 논문에서는 자동화된 시그니쳐 추출 시스템의 고려사항을 제시하고 새로운 응용의 인식에서부터 추출된 시그니쳐의 관리까지 시그니쳐 추출의 전 과정을 자동화하기 위한 시스템을 설계한다.

Ⅰ. 서 론

스마트 디바이스의 등장과 초고속 인터넷을 이용한 멀티미디어 응용 서비스가 대중화 됨에 따라 트래픽의 양이 급격히 증가하고 다양해지고 있다. 이에 따라 망 관리 비용이 증가하고 고품질 네트워크 관리 서비스 제공이 요구된다[1].

효과적인 망 관리를 위해서는 해당 트래픽을 발생시킨 원천(응용, 서비스, 프로토콜 등)을 판단하는 트래픽 분석이 선행되어야 한다. 분석 결과는 트래픽 제어 및 차단 등에 활용되기 때문에 정확성, 실시간성이 보장되어야 한다.

시그니쳐 기반 트래픽 분석 방법은 특정 응용이 발생시킨 트래픽에서 해당 응용을 구분할 수 있는 고유한 패턴을 시그니쳐로 추출하고 이를 이용하여 트래픽을 분석한다. 기존에 제안된 논문들은 시그니쳐 추출을 위해 수작업 및 일부 자동화된 방법을 제안하고 있다. 하지만, 시시각각으로 새로 출현하고 소멸되는 다양한 응용의 시그니쳐를 추출하기 위해서는 응용 인식에서부터 추출된 시그니쳐의 관리까지 자동화된 시스템이 필요하다.

시그니쳐 자동 추출을 위해 기존에 제안된 논문들은 오프라인으로 수집된 정답지 트래픽에서 시그니쳐를 수작업 없이 추출하는 방법에 초점을 맞추었다[2][3]. 하지만, 새로운 응용을 인식하거나 추출된 시그니쳐의 검증, 그리고 시그니쳐 관리에 대한 연구는 미비한 실정이다. 따라서 본 논문에서는 인터넷 응용 트래픽 분석을 위한 자동화된 시그니쳐 추출 시스템의 고려사항과 시스템 구조를 제시한다.

본 논문의 구성은 다음과 같다. 2 장에서는 제안하는 시스템의 고려사항에 대해 설명하고. 3 장에서는 자동화된 시그니쳐 추출 시스템을 제안한다. 마지막으로 4 장에서는 결론 및 향후 연구에 대해 기술한다.

Ⅱ. 고려사항

본 장에서는 자동화된 시그니쳐 추출 시스템을 개발할 때, 요구되는 고려사항들을 제시한다.

•다차원 분류 체계

명확한 분류 기준에 기반하지 않은 분석 결과는 활용성 측면에서 많은 제한 점을 가진다[4]. 예를 들어 "프로토콜"을 분류 기준으로 사용한 분석 결과는 "응용"단위의 트래픽 제어 및 차단에 활용되기 어렵다. 기존의 많은 연구에서는 명확한 분류 기준 없이 분류 방법에 의존적인 분류 기준을 혼합하여 사용함으로써, 정확한 분석을 어렵게 하였다. 따라서, 계층적인 다양한 분류 기준을 가지는 명확한 분류체계를 고려 해야 한다.

•다양한 추출 네트워크

네트워크의 종류와 구성원들의 인터넷 이용 목적에 따라 해당 네트워크에서 사용되는 응용은 차이가 있다. 따라서, 다양한 종류의 네트워크(회사, 학교, 가정)에서 발생하는 정답지 트래픽을 수집하여 시그니쳐를 추출한다. 즉, 다양한 네트워크에 설치된 추출 시스템의 시그니쳐를 통합하는 관리 기술이 고려 되야 한다.

•다양한 시그니쳐 모델

독점 프로토콜 사용, 암호화 등 응용 트래픽은 점점 복잡해지고 다양해 지고 있다. 기존의 단순한 헤더, 페이로드 기반의 시그니쳐 모델로는 특정 응용의 고유한 패턴을 기술하기에 역부족이다. 따라서 다양한 트래픽의 특징을 기술할 수 있는 다양한 시그니쳐 모델(통계, 행동기반 등)이 고려 되야 한다.

이 논문은 2010 년도 정부(교육과학기술부)의 재원으로 한국연구재단-차세대정보컴퓨팅기술개발사업의 지원을 받아 수행된 연구임(No.20100020728).

•**검증 네트워크**

추출된 시그니쳐의 검증은 시그니쳐의 분석 성능을 측정하는 작업이다 정확한 성능 평가를 위해서는 명확한 분류 기준과 해당 분류 기준에 맞는 평가 방법론이 필요하다. 또한, 특정한 목적에 국한된 평가가 아닌 다양한 측면(분석결과, 결과활용, 실시간성)의 평가가 고려 되야 한다.

Ⅲ. 자동화된 시그니쳐 추출 시스템

본 장에서는 자동화된 시그니쳐 추출 시스템을 보인다. 본 시스템은 응용 인식을 시작으로 시그니쳐 관리까지 총 5 가지 모듈로 구성되어 있다. 새로운 응용을 인식하고 해당 응용의 정답지 트래픽을 분류 체계 기반으로 수집한다. 수집된 정답지 트래픽에서 다양한 시그니쳐 모델을 기반으로 시그니쳐를 추출하고 추출된 시그니쳐를 다양한 관점의 성능 평가 기준을 적용하여 검증한다. 검증된 시그니쳐는 최적의 분석 성능을 유지하기 위해 관리된다. 본 시스템은 복수 네트워크에 설치되어 사람의 수작업 없이 자동으로 수행된다. 그림 1 은 자동화된 시그니쳐 추출 시스템의 구조를 보여준다.

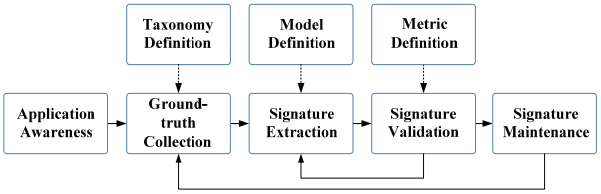

그림 1. 자동화된 시그니쳐 추출 시스템

Application Awareness(응용 인식) 응용의 생성과 소멸이 역동적으로 이루어지는 네트워크 환경에서 새로운 응용(시그니쳐 추출 대상)을 빠르게 인식해야 한다. 새로운 응용을 자동으로 인식하기 위해 Agent 기반(종단 호스트에서 응용정보 수집), User-Agent 필드 기반(HTTP 를 사용하는 응용정보 수집), DNS 응답 패킷 기반(웹사이트의 도메인 정보 수집)방법을 적용한다.

Taxonomy Definition(분류체계 정의) 불명확한 분류 체계를 기반한 트래픽 분석 결과는 부정확할 뿐만 아니라 활용성 측면에서도 많은 한계점을 가진다. 따라서 다각적이고 계층적인 다양한(서비스, 응용, 프로토콜, 기능, 콘텐츠) 분류기준을 정의한다.

Ground-truth Collection(정답지 트래픽 수집) 다차원 분류 체계의 분류 기준에 맞는 시그니쳐를 추출하기 위해 발생 원천이 명시된 정답지 트래픽을 각 분류 기준에 맞게 수집한다. 각 분류 기준에 맞는 정답지 수집 방법(서비스:DNS 분석, 응용:Agent 분석, 프로토콜:행동 분석, 기능:통계 분석, 콘텐츠:헤더 분석)을 적용한다.

Model Definition(모델 정의) 응용 시그니쳐는 트래픽의 원천을 구분할 수 있는 해당 응용의 고유한 특징이다. 따라서 시그니쳐는 트래픽의 특성을 잘 반영할 수 있고 서로 한계점을 상호 보완할 수 있는 다양한(페이로드, 헤더, 통계, 행동기반) 형태의 시그니쳐 모델을 정의한다.

- **페이로드:** 트래픽 페이로드에 존재하는 특정 패턴(바이트의 시퀀스)
- **헤더**: 고정된 서비스를 일정 기간 제공하는 서버의 헤더 정보
- **통계**: 플로우(패킷의 집합)에서 패킷의 크기, inter-arrival time, window size 등에 대한 통계적 정보
- **행동기반**: 응용이 트래픽을 발생할 때 관찰되는 독특한 플로우 간의 패턴

Signature Extraction(시그니쳐 추출) 다양한 형태의 시그니쳐 모델을 기반으로 정답지 트래픽에서 시그니쳐를 자동 추출한다. 정확하고 신속한 시그니쳐 추출을 위하여 모델에 최적화된 다양한 방법론(페이로드:longest common subsequence algorithm, 헤더:association rule mining, 통계: clustering algorithm, 행동기반:automata)을 적용한다.

Metric Definition(평가기준 정의) 추출된 시그니쳐의 정확한 성능 평가를 위한 객관적이고 다양한 관점의 평가 기준을 정의한다. 기존에 사용되던 분석률(분석 대상 트래픽 중에서 분석된 트래픽의 비율), 정확도(분석된 트래픽 중에서 정확히 분석된 트래픽의 비율)뿐만 아니라, 각 응용 별, 시그니쳐 별, 적용 환경 등 다양한 평가 기준을 제시한다.

Signature Validation(시그니쳐 검증) 평가 기준을 신속, 정확하게 측정 할 수 있는 평가 방법론을 개발하여 추출된 시그니쳐를 검증한다. 검증은 추출 대상 정답지 트래픽 뿐만 아니라 다른 응용의 정답지 트래픽에도 적용한다. 검증 결과, 일정 수준 미달의 성능을 가지는 시그니쳐는 시그니쳐 추출 모듈에 알람 신호를 보냄으로써 분석 성능이 우수한 시그니쳐를 추출하도록 한다.

Signature Maintenance(시그니쳐 관리) 인터넷 응용의 개수가 급격하게 증가하고 다양한 기능들이 추가, 수정되기 때문에 분류 성능 유지를 위한 시그니쳐 관리가 필요하다. 더 이상 사용되지 않거나 오탐율이 높은 시그니쳐를 폐기하고, 복수의 네트워크에서 다양한 모델을 기반으로 추출된 시그니쳐의 중복성을 고려하여 최적의 시그니쳐로 통합한다. 또한, 다양한 네트워크에 필요한 시그니쳐를 선택적으로 배포한다. 재 추출이 필요한 경우, 정답지 트래픽 수집 모듈에 알람 신호를 보내 재 추출한다.

Ⅳ. 결론 및 향후 연구

새로운 응용의 시그니쳐를 자동으로 추출하기 위해 자동화된 시그니쳐 추출 시스템을 설계하였다. 본 시스템은 새로운 응용의 인식부터 추출된 시그니쳐의 관리까지 모든 과정이 자동으로 수행된다. 또한, 시스템 개발 시 고려되어야 할 사항들에 대해 제시하였고, 시스템의 각 모듈의 기능에 대해 설명하였다.

향후 연구에서는 본 논문에서 제안한 시스템의 세부 모듈들을 구현하고 통합된 시스템으로 개발하겠다.

참 고 문 헌

[1] S. Sen and J. Wang. "Analyzing peer-to-peer traffic across large networks," 2002 ACM SIGCOMM Internet Measurement Workshop, Marseilles, France, Nov. 2002.

[2] B. Park, Y. J. Won, M. Kim, and J. W. Hong."Towards automated application signature generation for traffic identification," Proc.NOMS'08, pp. 160-167, 2008.

[3] MU Cheng1, HUANG Xiao-hong, TIAN Xu, MA Yan, Qi Jing-li. "Automatic traffic signature extraction based on fixed bit offset algorithm for traffic classification," The Journal of China Universities of Posts and Telecommunications, 18(2), pp. 79-85, Dec. 2011.

[4] Ji-hye Kim, Sung-Ho Yoon and Myung-Sup Kim, "Research on Traffic Taxonomy for Internet Traffic Classification," Proc. APNOMS'11, pp. 21-23, Sep. 2011.