트리 구조 기반 CSP(Contiguous Sequence Pattern) 알고리즘을 이용한 네트워크 비공개 프로토콜 구조 추론

심규석, 구영훈, 박지태, 김명섭

고려대학교

{kusuk007, gyh0808, pjj5846, tmskim}@korea.ac.kr

Inference of Network Non-Public Protocol Structure using CSP(Contiguous Sequence Pattern) Algorithm based on Tree Structure

Kyu-Seok Shim, Young-Hun Goo, Jee-Tae Park and Myung-Sup Kim

Korea Univ.

요 약

오늘날 네트워크 분야의 발전이 증진되면서, 대부분의 응용이 네트워크 기능을 사용한다. 기존에는 미리 정의된 형식의 프로토콜을 사용하여 네트워크 기능을 사용하였다면, 최근 해당 응용만을 위한 전문적인 프로토콜을 개발하여 보다 효율적인 방식으로 네트워크 기능을 사용하고 있다. 사용자와 개발자에게는 매우 효율적으로 응용에서 필요한 데이터를 송, 수신할 수 있지만, 네트워크 관리자 입장에서는 해당 트래픽을 관리하기 매우 복잡화되었다. 따라서, 본 논문에서는 이러한 비공개 프로토콜의 구조를 자동으로 추론하기 위해 트리 구조 기반의 CSP(Tree Contiguous Sequence Pattern)방법을 사용한다. 트리 구조 기반 CSP는 공통 문자열을 추출하기 위해 기존 CSP방법과 Tree 구조를 결합시켜 비공개 프로토콜 구조를 추론하는데 사용한다. 본 방법은 HTTP 프로토콜을 통해 비공개 프로토콜의 구조를 추론할 수 있는 가능성을 확인하였다.

Ⅰ. 서 론

오늘날 네트워크 분야의 발전은 크게 증진되었고, 더 크게 증진될 예정이다. 대부분의 응용은 네트워크 기능을 사용한다. 이러한 응용들은 기존 정의된 프로토콜이 아닌 자신만의 프로토콜을 이용하여 좀 더 효율적으로 네트워크 기능을 사용한다.

해당 응용에서 전문적으로 사용하는 프로토콜은 사용자와 응용 서비스 제공자 입장에서는 네트워크 자원을 빠르게 효율적으로 사용하는 유용한 방법이지만, 네트워크 관리자입장에서 응용만이 사용하는 프로토콜은 비공개 프로토콜로 구분되어 트래픽 분석 시 매우 어려움을 겪고 있다. 해당 프로토콜은 미리 정의되어 있지 않기 때문에 포맷, 구문, 구조 등에 대한 정보가 없어서 트래픽 분석하기 어렵다.

따라서 해당 프로토콜을 사용하는 트래픽을 분석하기 위해 프로토콜의 포맷, 구문을 추론할 수 있는 방법이 요구된다. 비공개 프로토콜의 구조를 추론해야 해당 프로토콜을 사용한 트래픽의 의미에 대해 자세한 분석이 가능하며, 보다 효율적인 네트워크 트래픽 관리를 위해 필수적이다.

본 논문에서는 이러한 프로토콜의 구조를 추론하기 위해 메시지 단위 분류와 TCSP(Tree Contiguous Sequence Pattern)을 사용한 구조 추론 방법을 사용한다. 기존 Apriori[2] 기반의 CSP[1]에 Tree를 접목시켜 새로운 방법으로 고정필드를 찾아낸다.

본 논문은 본장 서론에 이어, 2장 본론에서 전처리, CSP방법과 TCSP 방법에 대해 언급하고, 3장 결론으로 본 논문을 마친다.

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

Ⅱ. 본론

본 논문에서는 네트워크 비공개 프로토콜 구조를 추론하기 위한 TCSP 방법을 제안한다. 먼저, 비공개 프로토콜 구조를 추론하기 위해서 전처리 과정에서 네트워크 트래픽에 대한 분류가 필요하다. 해당 분류를 Message Assemble단계로 정의하고, Message Assemble 단계에서 트래픽에 대해 Request, Response로 나눈다. 나누어진 트래픽에 대해 메시지 단위를 정의하며 전처리 과정을 마친다.

다음단계에서 비공개 프로토콜의 구조를 추론하기 위해 프로토콜에서 사용되는 고정적인 구문을 추출한다. 고정적인 구문을 추출한 뒤 해당 메시지들에 대해 CSP를 이용하여 공통적으로 발생하는 구분에 대해 추출한다. 본 방법을 HTTP 프로토콜을 예제로 가능성을 제시한다.

Ⅱ-1. Message Assemble

Message Assemble 단계는 프로토콜의 고정 필드를 찾기위한 전처리 단계이다. 본 전처리 단계에서는 입력으로 들어온 단일 프로토콜에 대한 분류 작업을 진행한다. 메세지를 분류하는 것은 후에 사용될 TCSP가 Apriori 기반이기 때문에 해당 메시지가 Support의 단위가 된다. 따라서 메시지 분류에 따라 TCSP 결과가 달라질 수 있기 때문에 매우 중요한 전처리 작업이다.

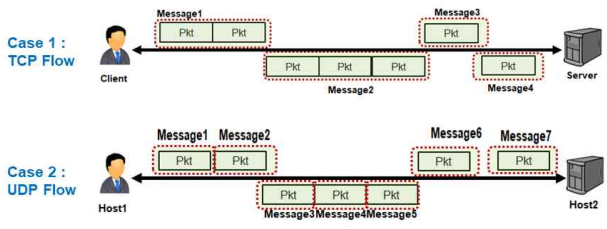

그림 1. Message Assemble 과정

본 단계에서 그림1에서와 같이 Message Assemble은 두가지 경우로 구분된다. 첫 번째는 TCP의 경우, 방향이 같은 연속된 패킷을 하나의 메시지로 정의한다. TCP는 대부분 하나의 의미있는 메시지 뒤로 전송해야하는 데이터 부분이 연속으로 전송되기 때문에 다음과 같이 메시지를 정의한다. 두 번째는 UDP의 경우, 각 패킷을 하나의 메시지로 정의한다.

Ⅱ-2. CSP (Contiguous Sequence Pattern)

CSP는 Apriori 기반으로 연속된 공통 문자열을 추출하는 방법이다. 단계는 다음 그림과 같이 진행된다. 두 개의 시퀀스가 존재할 때, 두 개의 시퀀스에서 공통 문자열을 찾기 위한 방법으로 다음과 같이 길이 1부터 최대 길이까지 Merge, Delete 과정을 반복하며 최종적으로 공통 문자열을 추출할 수 있다.

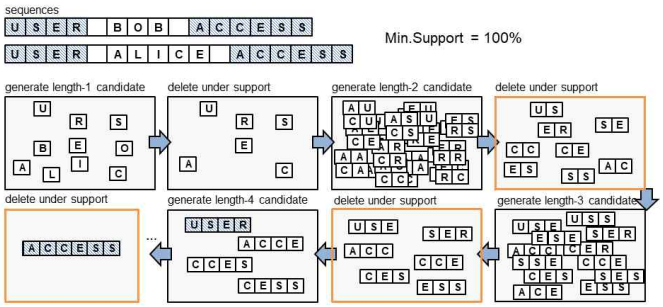

그림 2. CSP 과정

Ⅱ-3. 트리 구조 기반 CSP (Contiguous Sequence Pattern)

본 논문에서는 Ⅱ-2에서 언급한 CSP를 Tree형태로 이용하여 비공개 프로토콜의 구조를 추론한다. Ⅱ-1에서 분리된 메시지에서 CSP를 이용하여 가장 높은 Support를 가지는 최대길이 공통 문자열을 추출한다. 추출된 공통 문자열을 기준으로 앞부분과 뒷부분을 나누는 Binary Tree가 만들어진다. 앞부분은 다시 위의 과정을 더 이상 공통 문자열이 추출되지 않을 때까지 반복한다. 만약 추출된 공통 문자열의 Support 값이 100%가 아니라면 추출에 사용되지 않은 메시지들만을 가지고 다시 공통 문자열을 추출한다. 본 과정을 반복하게 되면, 같은 위치에서 발생하는 키워드들에 대해 모든 경우에서 키워드를 추출할 수 있다. 다음 그림은 TCSP 과정을 나타내기 위한 HTTP Request 트래픽을 대상으로 한 예시이다. 그림 3은 예시에 사용된 메시지 내용이고, 같은 색의 트래픽은 같은 종류의 트래픽을 의미한다.

그림 3. TCSP에서 사용되는 HTTP 메시지

다음 그림 4와 같이 위의 과정을 진행하면, HTTP 프로토콜에서 사용되는 모든 키워드를 추출할 수 있다. 기존 방법에서 POST의 경우 같은 위치에서 GET의 Support가 더 높기 때문에 GET은 추출될 수 있지만 POST는 추출되지 않는 문제를 본 방법에서 해결할 수 있다.

본 방법을 통해 고정필드를 추출한 뒤, 고정필드를 이용하여 모든 메시지 포맷을 구성한다. 해당 메시지 포맷들은 프로토콜에서 추출될 수 있는 모든 형태로 추출될 것이다.

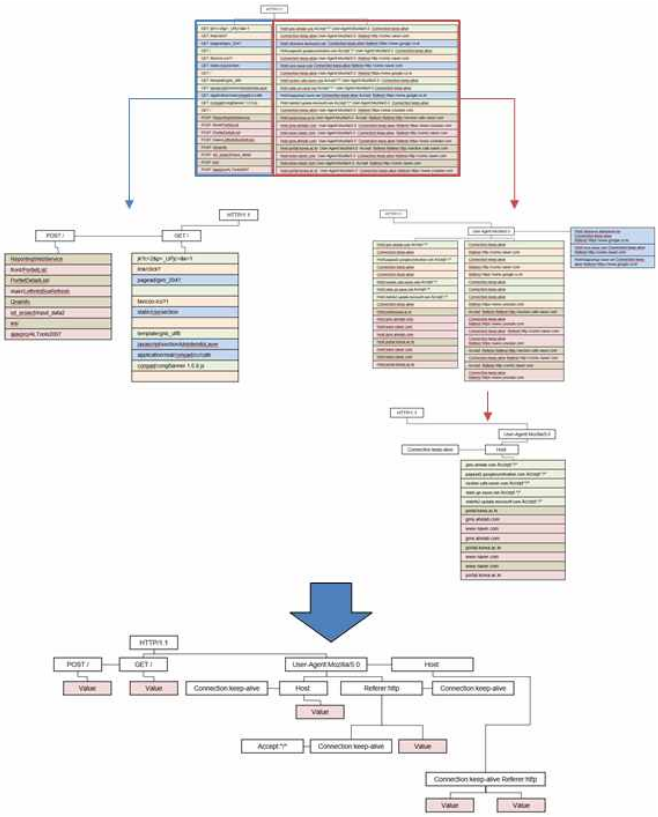

그림 4. 비공개 프로토콜 고정 필드를 추출하는 TCSP 과정

Ⅲ. 결론

본 논문에서는 정의된 프로토콜이 아닌 효율적인 데이터 전송을 위한 응용 만의 전문적인 프로토콜 즉, 비공개 프로토콜을 네트워크 관리자 입장에서 효과적으로 분석하기 위한 트리 구조 기반 CSP를 이용하여 프로토콜 구조 추론하는 방법은 제안한다. 본 방법은 기존 CSP만을 사용한 방법으로 추출되지 않았던 고정필드를 추출할 수 있는 결과를 확인할 수 있었고, 효과적으로 모든 고정 필드를 추출할 수 있는 가능성을 확인하였다.

향후 본 방법을 통해 HTTP 프로토콜 뿐만 아니라 DNS등 Binary로 구성된 프로토콜에서 가능성을 확인하고, 메시지 포맷을 확인하여 성능평가를 할 예정이다.

참 고 문 헌

[1] Kyu-Seok Shim, Young-Hoon Goo, Sungyun Kim, Mi-Jung Choi and Myung-Sup Kim, "SigManager: Automatic Payload Signature Management System for the Classification of Dynamically Changing Internet Applications," Proc. of the Asia-Pacific Network Operations and Management Symposium (APNOMS) 2017, Seoul, Korea, Sep. 27-29, 2017, pp.350-353.

[2] R. Agrawal and R. Srikant, "Mining sequential patterns," in Data Engineering, 1995. Proceedings of the Eleventh International Conference on, 1995, pp. 3-14.