프로토콜 구조 분석에서 CSP 알고리즘과 SP 알고리즘을 적용한 메시지 포맷 추출 성능 비교

구영훈, 심규석, 이민섭, 김명섭
고려대학교
{gyh0808, kusuk007, chenlima2, tmskim}@korea.ac.kr

Comparison of Message Format Extraction Performance Using CSP Algorithm and Using SP Algorithm in Protocol Structure Analysis

Young-Hoon Goo, Kyu-Seok Shim, Min-Seob Lee, Myung-Sup Kim
Korea Univ.

요 약

최근 급격한 인터넷의 발전은 인간 사회에 지대한 혜택을 가져다주었지만, 이에 따라 발생하는 대용량의 복잡한 트래픽은 네트워크의 관리를 더욱 어렵게 만들고 있으며 이는 꾸준한 보안사고의 발생과 국가 간 사이버 갈등의 심화로 나타나고 있다. 효율적인 네트워크의 관리 및 보안을 위해 비공개 프로토콜에 대한 구조분석 기술 확보가 시급한 시점이다. 본 논문에서는 최근 주목받고 있는 데이터마이닝 기술을 이용한 비공개 프로토콜의 메시지 포맷 추출 방법론 중 CSP(Contiguous Sequential Pattern) 알고리즘을 적용하였을 때와 일반적인 순차 패턴 알고리즘을 적용하였을 때의 추출된 메시지 포맷 성능을 실험을 통해 비교한다.

Ⅰ. 서론

최근 급격한 인터넷의 발전은 인간 사회에 지대한 혜택을 가져다주었지만, 이에 따라 발생하는 대용량의 복잡한 트래픽은 네트워크의 관리를 더욱 어렵게 만들고 있으며 이는 꾸준한 보안사고의 발생과 국가 간 사이버 갈등의 심화로 나타나고 있다. 또한 일상생활이 정보통신 기술 기반의 사이버 공간상에 의존하고 있는 오늘날, 사이버테러는 지구촌의 새로운 공동 관심사로 부상하고 있으며, 이에 사용되는 비공개 프로토콜에 대한 구조분석 기술 확보가 시급한 시점이다.

국내외적으로 비공개 프로토콜의 구조분석을 위한 다양한 프로토콜 리버스 엔지니어링 기술이 연구되고 있는 가운데, 최근 데이터마이닝 기술을 활용한 메시지 포맷 추출 방법론이 주목받고 있다[1-4]. 본 논문에서는 이러한 데이터마이닝 기술을 활용한 메시지 포맷 추출 방법론들의 크게 인접한 연속 순차 패턴 마이닝과 단순 순차 패턴 마이닝으로 구분하고 추출된 메시지 포맷의 성능을 비교한다. 두 알고리즘을 통해 추출된 메시지 포맷 성능을 비교함으로써 향후 데이터마이닝 기술을 이용한 메시지 포맷 추출 방법에 대한 연구 방향을 제시한다.

본 논문은 2장에서 관련 연구에 대해 설명하고, 3장에서는 메시지 포맷 추출을 위한 인접한 연속 순차 패턴 마이닝과 단순 순차 패턴 마이닝을 비교하여 기술하며 4장에서는 이 두 알고리즘을 통해 추출한 메시지 포맷의 성능을 실험을 통해 분석한다.

Ⅱ. 관련 연구

자동 프로토콜 리버스 엔지니어링은 크게 실행 트레이스 기반 방법론과 네트워크 트레이스 기반 방법론으로 구분된다. 이 중 네트워크 트레이스 기반 방법론은 오직 트래픽 트레이스만을 분석하며 프로토콜을 사용하는 엔티티에 대한 별도의 제어를 요구하지 않기에 보다 현실적으로 분석에 용이하므로 자주 적용되는 방법론이다. 네트워크 트레이스 기반 방법론은 생물정보학에 바탕을 둔 Sequence Alignment와 관련된 기술을 사용하거나 자연언어 처리에 바탕을 둔 빈도와 관련된 기술을 사용한다. 전자의 경우, 한 번에 오직 두 개의 메시지 시퀀스만을 입력으로 하여 정렬하기 때문에 여러 메시지 시퀀스를 정렬하는 데에 지수적 복잡도가 소요된다. 후자의 경우, 빈도에 기반을 둔 기술에 의존하므로 빈도와 관련한 특징이 없는 메시지 구조를 추론하는데 한계가 있다. 이 가운데 최근 주목받는 방법론은 순차 패턴 마이닝을 활용한 방법론이다. 순차 패턴 마이닝은 데이터에 공통으로 나타나는 순차적인 패턴을 찾아내는 것이다.

[1]은 순차 패턴 마이닝 중 Apriori 알고리즘을 사용하여 필드 포맷과 메시지 포맷을 추출한다. [2]와 [3]은 Aho-Corasick 알고리즘을 사용하여 필드 포맷을 추출한 후 순차 패턴 마이닝 중 FP-growth 알고리즘을 사용하여 메시지 포맷을 추출한다. FP-growth 알고리즘은 Apriori에서 제시한 후보의 생성에 많은 시간이 걸린다는 단점을 고려하여, 후보 생성을 하지 않고 자주 발생하는 항목집합 모두를 생성하는 것으로 시작한다. Apriori와 FP-growth의 성능을 비교 연구한 결과에서 길고 짧은 자주 발생하는 패턴들을 마이닝할 때 FP-growth 방법이 빠르며 균형적이다. [4]는 단순히 메시지에서 발생되는 시계열 서브 시퀀스를 추론하는 순차 패턴 마이닝이 아닌 인접한 연속 순차 패턴 마이닝에 AprioriTID와 AprioriHash를 통합한 CSP 알고리즘을 사용한다.

Ⅲ. CSP 알고리즘과 SP 알고리즘의 비교

본 논문에서는 인접한 연속 순차 패턴 마이닝을 CSP(Contiguous Sequential Pattern) 알고리즘이라 명명하고 단순 순차 패턴 마이닝을 SP(Sequential Pattern) 알고리즘이라 명명하여 두 알고리즘을 통해 추출한 메시지 포맷의 성능을 비교한다. 본 논문에서 사용하는 용어인 CSP는 [4]에서 사용하는 CSP가 아니며 이를 포괄하는 더 큰 개념을 의미한다. SP 알고리즘은 인접한 연속 순차 패턴이 아닌 단순히 시계열에 따른 항목들의 시퀀스를 추출하는 알고리즘을 의미한다. 메시지 포맷을 추출하는 방법으로 [4]는 CSP를 사용하고 있으며 [1-3]은 명확한 언급은 없으나 필드 포맷 추출 방법론과 메시지 포맷 추출 결과를 보아 CSP를 사용하는

이 논문은 2018년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2018R1D1A1B07045742).

것으로 추론할 수 있다. 필드 포맷을 추출하기 위해서는 연속 바이트 스트림에 대한 키워드를 추론하여야하므로 CSP 알고리즘을 사용하여 추출하여야 한다. 그러나 메시지 포맷을 추출할 때에는 꼭 CSP 알고리즘이 아닌 SP 알고리즘을 사용하는 것을 고려할 수 있다.

CSP 알고리즘을 사용하여 메시지 포맷을 추출할 시 항상 인접하여 연결되어 있는 순차 패턴만을 추출하므로 다양한 메시지 포맷의 추출이 불가능하며 특히 HTTP의 Header Line이나 DNS의 Query, Answer, Authoritative, Additional Section과 같이 항상 필드의 위치와 순서가 가변적이며 선택적으로 존재하는 레코드 타입의 필드가 존재하는 메시지 구조의 경우 메시지의 처음부터 끝까지 표현하는 메시지 포맷이 아닌 일부분만을 표현하는 단편적인 메시지 포맷만을 추출하게 된다. 그러나 인접하지 않은 모든 시계열 순차 패턴을 추출하는 SP 알고리즘을 사용할 경우 각 Level에서 더 다양한 패턴을 추출할 수 있으므로 더 큰 Length의 패턴을 추출하는 Level로의 진입을 기대할 수 있어 메시지 포맷의 세분화를 가능케 한다. 이를 통해 레코드 타입의 필드가 존재하는 메시지 구조 또한 더 상세하게 추출이 가능하다.

Ⅳ. 실험 및 결과

CSP 알고리즘과 SP 알고리즘을 적용한 메시지 포맷의 추출 성능을 비교하기 위해 [5]에서 제시한 메시지 포맷의 세분화 정도를 의미하는 Detail 성능 평가 지표를 활용한다. HTTP 프로토콜과 DNS 프로토콜에 대하여 각각 CSP 알고리즘과 SP 알고리즘을 적용하여 메시지 포맷을 추출하였다. 그림 1은 HTTP 메시지 포맷들의 개별 Detail 수치에 대한 히스토그램과 정규 확률 분포 그래프이다. CSP를 사용하였을 경우, 추출된 메시지 포맷은 36개이며 개별 Detail의 최대값, 중앙값, 평균은 각각 133%, 28.8%, 40.6%의 결과를 보였다. SP를 사용하였을 경우, 추출된 메시지 포맷은 528개이며 개별 Detail의 최대값, 중앙값, 평균은 각각 133%, 70.7%, 68.5%의 결과를 보였다. 결과적으로 전체 메시지 포맷의 세분화 정도는 SP를 사용하였을 때 약 27.9%가 더 향상되었으며, 그림 1의 정규 확률 분포 그래프에서 CSP보다 SP의 경우가 더 높은 Detail 계급 구간에 확률 분포가 치우쳐져 있음을 확인할 수 있다.

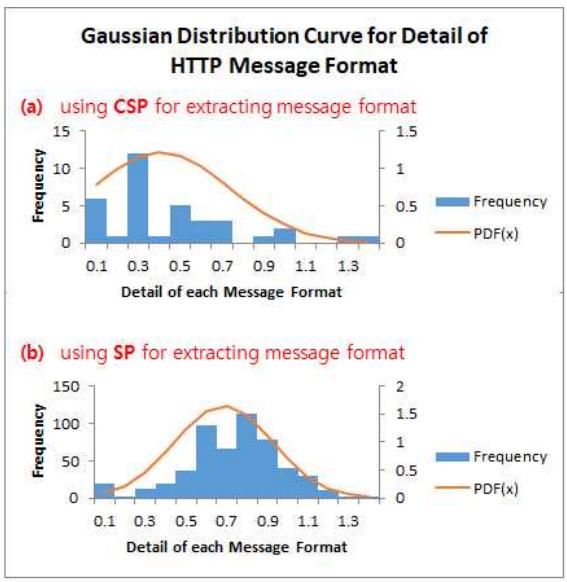

그림 1. CSP와 SP로 추출한 HTTP 메시지 포맷들의 Detail 정규 확률 분포

그림 2는 DNS 메시지 포맷들의 필드 포맷 개수에 대한 히스토그램과 정규 확률 분포 그래프이다. CSP를 사용하였을 경우, 추출된 메시지 포맷은 27개이며 내부 필드 포맷 개수의 최대값, 중앙값, 평균은 각각 9, 4, 4.4의 결과를 보였다. SP를 사용하였을 경우, 추출된 메시지 포맷은 564개이며 필드 포맷 개수의 최대값, 중앙값, 평균은 각각 18, 10, 10.5의 결과를 보였다. 더 복잡한 레코드 타입의 필드를 갖는 메시지 구조를 갖는 DNS 프로토콜의 경우 HTTP 프로토콜의 경우보다 전체적으로 필드의 개수가 더 크게 향상되었으며 메시지 포맷의 최대 필드 포맷 개수도 2배나 증가하였다.

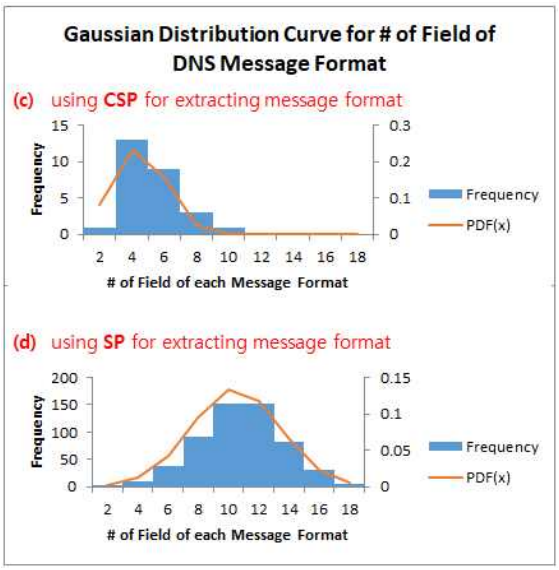

그림 2. CSP와 SP로 추출한 DNS 메시지 포맷들의 필드 포맷 개수 정규 확률 분포

결과적으로 SP 알고리즘을 사용하여 메시지 포맷을 추출하였을 경우 CSP 알고리즘을 사용하여 메시지 포맷을 추출하였을 때보다 [5]의 Correctness와 Coverage는 유지한 채 메시지의 세분화 정도는 향상시키며 레코드 타입의 필드를 갖는 메시지 구조 추출 또한 어느 정도 해결할 수 있음을 보였다. 그러나 Compression이 낮아지는 문제가 있으므로, 추출된 메시지 포맷의 우선순위에 따른 필터링에 관한 연구가 필요하다.

참 고 문 헌

[1] J.-Z. Luo, and S.-Z. Yu. “Position-based automatic reverse engineering of network protocols”, Journal of Network and Computer Applications, Vol.36, No.3, Issue.3, pp.1070-1077, Feb. 2013.

[2] Y. Wang, N. Zhang, Y.-M. Wu, B.-B. Su, and Y.-J. Liao, “Protocol Formats Reverse Engineering Based on Association Rules in Wireless Environment”, in Proceedings of the 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communication (TrustCom ‘13), pp.134-141, Melbourn, Australia, July. 2013

[3] R. Ji, H. Li, and C. Tang, “Extracting keywords of UAVs Wireless Communication Protocols Based on Association Rules Learning”, in Proceedings of the 12th IEEE International Conference on Computational Intelligence and Security, pp.310-313, Wuxi, China, Dec. 2016.

[4] Y.-H. Goo, K.-S. Shim, J.-T. Park, B.-M. Chae, H.-W. Moon, and M.-S. Kim, “A Method of Protocol Reverse Engineering for Clear Protocol Specification Extraction”, KNOM Review, Vol.20, No.2, pp.11-23, Dec. 2017.

[5] Y.-H. Goo, K.-S. Shim, M.-S. Lee, J.-T. Park, and M.-S. Kim, “A Study of Evaluation and Validation Method for Protocol Reverse Engineering”, in Proceeding KICS 2017 Summer Conference pp.11-23, Dec. 2017.