Industrial Control System (ICS)에서 사용하는 비공개 프로토콜 구조 추론을 위한 메시지 클러스터링 알고리즘 성능 비교

심규석 [1], 이원혁 [1], 손일권 [1], 이은주 [1], 김명섭 [2]
한국과학기술정보연구원 [1], 고려대학교 [2]

{kusuk007, livezone, d2estiny, saranha}@kisti.re.kr[1], tmskim@korea.ac.kr[2]

The Comparison of Message Clustering Algorithm Performance for Unknown Protocol Reverse Engineering using Industrial Control System (ICS)

Kyu-Seok Shim[1], Wonhyuk Lee[1], Ilkwon Sohn[1], Eunjoo Lee[1] and Myung-Sup Kim [2]
KISTI[1], Korea Univ[2].

요 약

최근 산업현장의 자동화는 네트워크 분야와의 융합으로 급속도로 진행되고 있다. 산업기기들은 Industrial Control System (ICS)과 연결되어 설계된 명령을 전송하고, 로그를 기록하며 기존 사람이 직접 작업하던 환경과 달라지고 있다. 하지만 ICS에서 사용하는 프로토콜은 대부분 효율적인 데이터 전송을 위해 ICS를 생산한 회사에서 직접 개발하여 사용되고 있다. 따라서 ICS 환경의 네트워크 트래픽을 분석하는 것은 많은 한계가 존재하고, 네트워크 트래픽을 통한 산업현장의 모니터링을 위해서는 또다시 값비싼 장비를 구매할 수밖에 없다. 이러한 환경에서 네트워크 트래픽 분석을 통한 산업공정 모니터링 및 네트워크를 통한 악성행위 탐지를 위해 산업용 프로토콜 메시지 클러스터링 알고리즘 성능을 비교한다. 본 논문에서는 UPGMA, Mean-Shift, 그리고 K-means 알고리즘을 선정하여 각 알고리즘에 대한 결과를 비교한다.

Ⅰ. 서 론

최근 산업현장 자동화는 네트워크 분야와의 융합으로 급속도로 진전하고 있다. 산업 자동화는 기존 사람이 산업기기들을 조작하고, 명령하던 것을 네트워크를 통해 설계된 내용대로 산업기기들이 동작하는 것을 의미한다. 각 산업현장의 기기와 Industrial Control System (ICS)장비 및 Engineering Workstation (EWS)는 네트워크로 연결되어 EWS 를 통해 사용자가 설계한 내용을 ICS 장비로 전송하고, 받은 내용대로 ICS 장비는 각 산업기기에게 명령한다. 즉, 모든 산업공정과정이 네트워크로 연결되어 있고, 네트워크를 통해 데이터를 주고받는다[1].

해당 네트워크에서 사용되는 프로토콜은 보통 ICS 제조회사에서 자체 개발한 프로토콜을 사용한다. 기존 프로토콜을 사용하지 않고 자체적으로 개발한 프로토콜을 사용하는 이유는 효과적인 데이터 전송을 위해서이다. 기존 프로토콜을 사용하게 되면 불필요한 필드에 데이터를 넣어야 하는 경우가 발생하고, 해당 경우에 따라 데이터 전송률 및 전송속도가 늦어지게 된다. 따라서 데이터 전송 목적에 적합한 프로토콜을 개발함으로써 불필요한 데이터 전송을 예방한다.

하지만, 이러한 자체적으로 개발된 프로토콜은 대부분 보안 및 지적소유권에 의해 비공개 되어 있다. 프로토콜의 스펙이 공개되지 않은 트래픽을 분석하는 것은 매우 많은 어려움이 있다. 비공개 되어있는 프로토콜의 스펙을 분석하는 것은 민감한 부분일 수 있지만 본 논문에서 제시하는 클러스터링 방법을 통한 메시지 타입 분류는 정확한 스펙을 분석하는데 목적이 있는 것보다는 해당 프로토콜에서 발생되는 메시지 타입을 분류하는 것만으로도 산업용 네트워크에서 네트워크를 통한 악성행위 탐지 및 모니터링에 필요한 정보를 수집하는데 목적이 있다. 따라서 네트워크 트래픽을 분석하기 위해서는 프로토콜 스펙 분석 즉, 프로토콜 구조를 추론하는 것이 선행되어야 한다.

본 논문에서는 ICS 환경에서 사용되는 프로토콜의 구조를 추론하기 위해 프로토콜에서 발생하는 메시지의 타입을 구분하는 클러스터링 알고리즘들을 선정하여 비교 평가한다. ICS 환경에서 가장 적합한 클러스터링 알고리즘을 선정하기 위해 비교하는 알고리즘들은 mean-shift 알고리즘, 대표적인 PRE 오픈소스인 Netzob 에서 사용하는 UPGMA 알고리즘, 대표적인 클러스터링 알고리즘은 K-means 알고리즘이다[2].

본 논문은 본 장 서론에 이어, 각 알고리즘의 소개 및 실제 트래픽 비교 결과를 포함한 본론, 그리고 향후연구 및 논문을 정리하는 결론으로 구성된다.

Ⅱ. 본론

ICS 환경의 프로토콜 메시지 타입을 분류하기 위해 본 논문에서는 Schneider, GE pac 장비에서 직접 네트워크 트래픽을 수집하고, 수집된 트래픽을 메시지형태로

변환하는 전처리과정을 수행한다. 전처리과정은 Pcap 형태로 수집된 트래픽을 Packet 그리고 Flow with Packet 형태로 변환하고, 변환된 Flow with packet 에서 헤더정보와 페이로드 내용을 이용하여 메시지 형태로 추출한다. 메시지는 메시지 ID, 방향, 사이즈, 타입번호, 위치, 내용으로 구성된다.

먼저, 메시지 형태로 추출한 산업용 프로토콜 네트워크 트래픽을 클러스터링 알고리즘으로 타입 별 분류하기 위해 사이즈별로 분류해야 한다. 산업용 프로토콜의 특성상 가변길이 필드 즉, 길이가 변하는 필드가 다수 포함되어 있지 않기 때문에 1 차적인 메시지 타입 분류의 목적이 있고, 다른 하나의 목적은 정확한 클러스터링 알고리즘 수행을 위해서이다. 길이가 다른 메시지들을 분류하기 위해 클러스터링 알고리즘을 수행한다면 패딩과 같은 전처리작업을 해야 한다. 그러한 전처리작업을 통해 정확도는 낮아질 수밖에 없다. 따라서 메시지 길이 별 분류를 해야 한다.

정확한 클러스터링 알고리즘 평가를 위해서는 정답지를 만들어야 한다. 본 실험에서는 직접 눈으로 확인하여 메시지 타입이 같은 메시지들을 분류한다. 직접 분류하였을 때 수집된 메시지타입은 다음 표 1 과 같다. 메시지 타입은 Request 메시지와 Response 메시지를 구분한다. 즉, Connection 기능을 사용할 때 Modbus/TCP 트래픽의 Request 메시지 타입 개수는 19 개, Response 메시지 타입 개수는 13 개이다[3].

Table 1. The information of messages and message types

Function		Modbus/TCP	Ethernet/IP
Connect	# Messages	3,506	299
	# Message Types	32(19+ 13)	29(16+ 13)
EWS ->PLC	# Messages	7,532	938
	# Message Types	66(22+ 44)	28(15+ 13)
PLC ->EWS	# Messages	4,523	674
	# Message Types	74(63+ 11)	19(11+ 8)

클러스터링 알고리즘 중 K-means 와 UPGMA 는 사용자의 개입이 필요하다. 즉, 사용자가 직접 threshold 를 입력해야 하는데, 본 실험에서 사용자는 메시지들이 몇 개의 타입으로 분류되는지 사전에 알지 못하기 때문에 elbow method 를 사용한다. Elbow method 는 가장 적절한 threshold 를 계산하여 입력하는 방법이다. 따라서 K-means 와 UPGMA 는 항상 같은 군집의 개수를 추출해내기 때문에 실험결과에서는 K-means 만 평가한다. 그러나, 두 알고리즘의 수행방법은 차이가 난다. UPGMA 의 경우 단순거리계산이 아닌 많은 컴퓨팅 자원을 사용하기 때문에 데이터양이 많고, 신속하게 결과를 추출해야 하는 본 방법과 적합하지 않다. 하지만 mean-shift 알고리즘은 Circle 의 범위만 지정하면 군집의 개수를 자동으로 분류한다. 다음은 실험결과이다.

다음 표 1 에서와 같이 정답지와 유사하게 메시지 타입을 분류하는 알고리즘은 Mean-shift 알고리즘이고, 정답지보다 세부적으로 분류하는 알고리즘은 K-means 알고리즘이다. 본 실험결과를 통해 어떤 알고리즘의 성능이 우수하다는 결론 보다는 목적에 따라 최적의 클러스터링 알고리즘을 사용할 수 있는 사전정보를 제공할 수 있다. 예를 들면, 산업용 프로토콜의 메시지 타입이 개수를 추출하기 위해서는 mean-shift 알고리즘이 적절하고, 메시지 타입 분류 후 추가적인 정보를 추출하기 위해서는 세부적으로 분류하는 K-means 알고리즘을 선정하는 것이 적절한다.

Table 2. The result of clustering message types

Function	Clustering	Modbus/TCP	Ethernet/IP
Connect	K-means	76(27+ 49)	29(16+ 13)
	Mean-Shift	43(27+ 16)	29(16+ 13)
EWS ->PLC	K-means	83(36+ 47)	33(17+ 16)
	Mean-Shift	65(27+ 38)	27(14+ 13)
PLC ->EWS	K-means	44(14+ 30)	29(14+ 15)
	Mean-Shift	53(43+ 10)	25(13+ 12)

Ⅲ. 결론

본논문에서는 비공개 되어 있는 산업용 프로토콜의 트래픽 분석을 위해 프로토콜 구조 추론을 위한 메시지 클러스터링 알고리즘의 성능을 비교하였다. 총 3 가지 알고리즘인 K-means, UPGMA, 그리고 mean-shift 알고리즘을 비교하였다. K-means 와 UPGMA 의 경우 threshold 로 몇개의 군집으로 분류할 것인지 입력해야 하는 메커니즘을 가지고 있기 때문에 elbow method 를 이용하였고, 따라서 두개의 군집개수는 같아질 수밖에 없었다. 하지만, 컴퓨팅 자원을 과도하게 사용하는 UPGMA 의 메커니즘으로 인해 K-means 가 빠른 결과를 추출할 수 있었다. Mean-shift 의 경우 정답지 군집개수와 가장 유사하게 클러스터링 결과를 도출할 수 있었으며, K-means 의 경우 정답지보다 세분화된 군집의 결과를 도출하였다.

ACKNOWLEDGMENT

본 연구는 2020 년도 한국과학기술정보연구원(KISTI) 주요 사업 과제로 수행한 것입니다

참 고 문 헌

[1] Stouffer, Keith, Joe Falco, and Karen Scarfone. "Guide to industrial control systems (ICS) security." NIST special publication 800.82: pp 16-16. 2011

[2] G. Bossert, ‘‘Exploiting semantic for the automatic reverse engineering of communication protocols,’’ Ph.D. dissertation, Univ. Gif-sur-Yvette, Rennes, France, Dec. 2014.

[3] Goldenberg, Niv, and Avishai Wool. "Accurate modeling of Modbus/TCP for intrusion detection in SCADA systems." International Journal of Critical Infrastructure Protection 6.2 (2013): 63-75