A Lightweight Software Solution for Log Analysis

1. 서론

1.1. 연구 배경 및 목표

많은 GUI 기반 로그 분석 소프트웨어는 실시간 분석을 보장하지 못하거나, 웹 로그에 대한 백업 기능을 제공하지 못한다. 또한 설치 과정이 복잡하고 인터페이스가 사용자 친화적이지 않은 경우가 많다. 즉, '쉽게 설치하여 사용 가능한 경량화 된 범용 로그 해석 도구'개발이 필요한 상황이다.

선행 연구로 "A Lightweight Software Solution for Log Analysis"[1]을 진행했다. 해당 연구에서 개발한 시스템에는 몇 가지 부족한 점이 있다. 첫째로, '경량화'의 근거가 모호하다. 무엇에 비교하여 경량한지를 정량적으로 드러내지 않는다. 둘째로, DDoS와 같은 휴리스틱에 기반하여 방어해야 하는 공격을 감지하지 못한다. 셋째로, LLM을 사용하여 출력하는 보안 보고서의 할루시네이션 문제를 해결하지 못했다. 넷째로, 공격 감지에 사용되는 시그니처로서의 정규식 정의의 책임을 사용자에게 전가하는 것으로, 실제 시스템의 사용자가 될 '소규모 사업체 또는 개인'의 상황을 고려하지 못했다. 다섯째로, 실시간 로그 분류에 BERT 등의 언어모델 도입을 후속 연구로 미뤄 두었다. 이에 따라 상기한 문제를 해결할 후속 연구를 진행한다.

후속 연구의 목표는, "개인 사용자도 쉽게 설치할 수 있는 경량화 된 SW"라는 기존의 목표를 유지하면서 시스템의 공격 감지 정확도와 감지 가능한 공격 유형을 확대하는 것이다.

1.2. 연구 범위

본 연구는 서론에서 제시한 'CLEAR'의 핵심 문제를 해결하는 것에 초점을 맞춘다. 구체적으로는, 대량의 로그에서 특이점을 검출하여 LLM으로 보안 보고서를 작성하도록 하는 방법론의 개발과, CISC2010[1]에 기반하여 학습시킨 ALBERT 모델을 CLEAR에 도입하여 공격 감지의 정확도를 향상하는 것을 포함한다. 이때 정밀한 시그니처로서의 정규식 개발은 연구 범위가 아니며, 현 연구에서는 LLM과 ALBERT의 동작 부하를 줄이기 위한 도구로서, 정규식 기반 로그 검사의 중요성을 축소할 예정이다.

2. 기대 효과

본 연구는 소규모 사업체에서 웹 기반 시스템을 안전하게 운용할 수 있도록 돕는 간편하고

[1] CSIC2010: 스페인 국립연구소에서 생성한, 웹 공격 방어 시스템의 테스트를 목적으로 하는 데이터셋.

실용적인 도구를 제공한다.

본 연구는 웹 로그가 정상 접속인지 판별하는 새로운 알고리즘을 제안하고, 그것의 유효성을 검증한다.

3. 연구 추진 계획

3.1. 연구 내용 및 순서

연구는 다음과 같은 순서로 진행된다.

- **ALBERT 모델 훈련과 검증**: CISC2010 데이터셋을 훈련 데이터와 검증 데이터로 나누어, ALBERT모델을 선행 연구[2]에서 얻어낸 90.9% 이상의 정확도를 갖추도록 훈련한다.
- **정상 접속 판별 알고리즘 개발**: 기존의 시스템은 정규식 세트에 기반하여 '의심되는 로그'를 찾았으나, 본 연구에서는 LLM과 ALBERT의 사용을 최소화 하기 위하여 '정상 접속 로그'를 판별해야 한다. 이때 의심 로그를 정상 로그로 판별하는 오탐은 매우 치명적이므로, 충분히 낮은 오탐률을 가질 때까지 알고리즘 개선과 CISC 데이터셋을 통한 오탐률 계측을 반복한다.
- **ALBERT 모델 병합**: 정규식 세트로 수행하던 개별 로그에 대한 실시간 감지를 훈련된 ALBERT 모델로 교체한다. 정규식에 의한 시그니처 검사에 통과한 로그만 ALBERT에 의해 판단된다.
- **LLM 보안 보고서 작성 알고리즘 개선**: 기존에는 전체 로그를 LLM에 통째로 입력하는 원시적인 방법을 사용한다. 본 연구에서는 정규식에 의해 정상 접속을 걸러내고, 위험 로그를 소규모로 LLM에 입력하여 상향식으로 보안 보고서를 작성하는 파이프라인을 구축한다.
- **DDoS 감지 기능 추가**: 선행 연구[3]에서 제시한 DDoS 감지 알고리즘을, CLEAR의 실시간 로그 감지 시스템에 포함시킨다.
- **시스템 검증**: '정상 접속 판별 알고리즘'에 의한 오탐률 상승과 시스템 자원 소비량 하락을 정량적으로 계측하여 시스템을 검증하고 미세조정 하는 것으로, 적절한 트레이드오프를 성사시킨다.

3.2. 연구 일정표

단계	일시	성과물
1단계: ALBERT 모델 훈련 및 검증.	09/01 – 09/14	90.9% 이상의 정탐률을 확보한 ALBERT 모델.

2단계: 정상 접속 판별 알고리즘 개발.	09/14 – 09/21	의심 로그를 정상 로그로 판별하는 확률이 0에 수렴하는 정규식 세트.
3단계: ALBERT 모델 병합.	09/21 – 09/24	1단계와 2단계의 성과물을 조합하여, CLEAR에 병합 완료.
4단계: LLM 보안 보고서 작성 알고리즘 개선.	09/24 – 09/28	로그를 특정 단위로 분할하여, 작은 단위의 보안 보고서를 상향식으로 수집하는 파이프라인 확보.
5단계: DDoS 감지 기능 추가.	09/28 – 09/30	실시간 감지 시스템에 DDoS 감지 시스템 병합 완료.
6단계: 시스템 검증.	10/01 – 10/05	유효성 검증이 완료된 시스템 확보.

참고문헌

[1] 이시헌 외, “A Lightweight Software Solution for Log Analysis”, 2025.

[2] 조영복 외, "웹 방화벽 로그 분석을 통한 공격 분류: AutoML, CNN, RNN, ALBERT", Journal of The Korea Institute of Information Security & Cryptology, Vol.34, No.4, Aug, 2024.

[3] 박재연 외, “리눅스 아파치 웹 서버 실시간 로그 분석을 통한 공격 탐지 프로그램 개발”, 정보과학회 컴퓨팅의 실제 논문지, Vol.24, no.4, pp. 190-197, 2018.