클라우드 환경에서 자연어처리 기법을 활용한 취약점 분석 시스템 설계

송진수*, 이필원*, 신용태**
*숭실대학교 컴퓨터학과
**숭실대학교 컴퓨터학부
iko153@soongsil.ac.kr, pwlee@soongsil.ac.kr, shin@ssu.ac.kr

Designing a vulnerability analysis system using natural language processing techniques in a cloud environment

Jin-Su Song*, Pil-Won Lee*, Young-Tea Shin**
*Dept. of Computer Science, Soong-Sil University
**Dept. of Computer Engineering, Soong-Sil University

요 약

최근 4차 산업혁명의 기술이 발전하며 인공지능과 클라우드 컴퓨팅의 융합에 대한 연구가 활발하게 진행되고 있으며 클라우드 컴퓨팅에 컨테이너 기술을 접목한 새로운 컴퓨팅 환경이 주목받고 있다. 그러나 현재 사용되고 있는 컨테이너 기반의 가상화 기술은 컨테이너 실행에 필요한 파일과 설정 값을 포함하고 있는 컨테이너 이미지를 통해 배포하는 방식을 사용하고 다수의 컨테이너가 하나의 커널을 공유하기 때문에 취약한 패키지를 사용하는 컨테이너 이미지가 다수의 사용자와 공유 되어 시스템 보안이 매우 취약하다 이에 본 논문에서는 자연어처리 기법을 활용한 취약점 분석 시스템을 통해 컨테이너를 실행에 필요한 파일과 설정 값을 포함하고 있는 컨테이너 이미지에서 취약점을 분석하는 시스템을 제안한다.

1. 서론

최근 4차 산업혁명의 기술이 발전하며 인공지능과 클라우드 컴퓨팅의 융합에 대한 연구가 활발하게 진행되고 있으며 클라우드 컴퓨팅에 컨테이너 기술을 접목한 새로운 컴퓨팅 환경이 주목받고 있다. 컨테이너 기반 컴퓨팅 가상화 기술은 OS를 활용하여 프로세스를 격리하고 해당 프로세스가 액세스할 수 있는 CPU, 메모리 및 디스크의 양을 제어하는 운영체제 가상화의 형식을 사용한다. 컨테이너 기반의 가상화는 OS 커널을 공유함으로써 애플리케이션 마다 전체 OS 인스턴스가 필요하지 않아 경량화, 플랫폼 독립성과 이식성이 높아 기존의 가상화에서 하이버바이저를 활용하여 물리적 하드웨어를 가상화하는 방식을 대체하면서 각광받고 있다[1]. 그러나 현재 사용되고 있는 대부분의 컨테이너 기반 클라우드 플랫폼은 오픈소스로 구성되어 있어 보안에 취약하며 매년 신규 취약점이 급속도로 증가하고 있다. 컨테이너 기반의 가상화 기술은 컨테이너 실행에 필요한 파일과 설정 값을 포함하고 있는 컨테이너 이미지를 통해 배포하는 방식을 사용하고 다수의 컨테이너가 하나의 커널을 공유하기 때문에 취약한 패키지를 사용하는 컨테이너 이미지가 다수의 사용자와 공유 되어 시스템 보안이 매우 취약하다[2].

이에 본 논문에서는 자연어처리 기법을 활용한 취약점 분석 시스템을 통해 컨테이너를 실행에 필요한 파일과 설정 값을 포함하고 있는 컨테이너 이미지에서 취약점을 분석하는 시스템을 제안한다.
제안하는 시스템은 자연어 처리 기법을 활용하여 쿠버네티스 클라우드를 구성할 때 작성되는 이미지 파일의 데이터 중 취약점을 유발하는 값을 분석하여 제공한다.

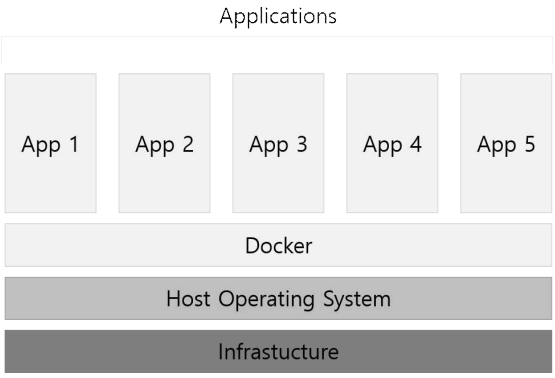
본 논문의 구성은 다음과 같다. 2장에서는 기존에 컨테이너 기반으로 사용 중인 클라우드 가상화 도구인 도커와 자연어 처리 기법에 대해 살펴본다. 3장에서는 본 논문에서 제안하는 클라우드 환경에서 자연어 처리 기법을 활용한 취약점 분석 시스템을 제시하고 마지막 4장에서는 결론 및 향후 연구 과제를

제시한다.

2. 배경 및 관련 연구

2.1 도커

도커(Docker)는 컨테이너 기반의 오픈소스 가상화 플랫폼으로써 다양한 프로그램과 실행환경을 포함하고 있는 이미지를 빌드하여 컨테이너로 추상화하고 동일한 인터페이스를 제공하여 개발자들의 프로그램의 배포 및 관리를 단순하게 해준다. [Fig. 1]은 도커의 구조도이다.



[Fig. 1] The structure of the docker

Docker는 오픈소스로 이루어진 가상화 도구로 리눅스 커널에서 컨테이너 방식으로 프로세스를 격리해서 실행한다. 각각의 컨테이너 안에 생성된 가상 운영체제는 호스트 운영체제에 있는 커널을 공유하기 때문에 컨테이너는 그자체가 호스트 운영체제의 자원을 일부 할당하여 작업을 수행하는 서버형태의 단일 프로세스로써 호스트 운영체제로부터 단순히 프로세스를 격리시키기 때문에 기존의 가상화 방식에 비해 가볍고 빠르게 동작하며 이를 통해 어플리케이션의 개발, 테스트, 서비스 환경을 하나로 통일하여 효율적으로 관리할 수 있다[2,3].

2.2 도커파일

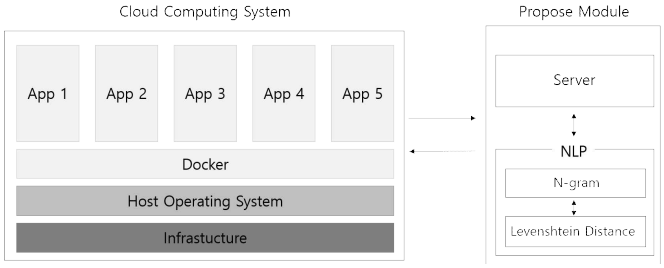
도커파일은 컨테이너를 생성하기 위한 이미지이고 컨테이너는 이미지가 실제 메모리에 로딩된 상태로 컨테이너를 실행하기 위한 모든 정보를 포함한다. 도커 클라우드를 구성하는 컨테이너를 배포하기 위해 컨테이너 명세서인 도커파일을 작성한다. .도커파일은 이미지를 생성할 때의 필요한 명령어의 집합으로 정의할 수 있으며, 설치해야하는 패키지, 소스코드, 컨테이너 구동과 동시에 실행 되어야하는 명령어와 셸 스크립트 등을 순차적으로 실행하여 사용자가 원하는 이미지를 만들 수 있도록 파일형태로 작성한다[3].

2.3 자연어 처리 기법

n-gram 언어 모델은 중복 단어의 빈도수를 기반으로 분석하는 통계적 접근방식으로 SLM의 하나이다. 자연어로 표현된 텍스트의 특징을 조사하기 위해 동일한 부분의 텍스트가 반복되는지 확인하는데 이때 n개씩 잘라낸 텍스트 중 같은 텍스트가 발견되면 카운트하여 분석하는 방식으로 문장의 유사도를 확인할 수 있다. n-gram의 종류는 n이 1일 때는 유니그램, 2일 때는 바이그램, 3일 때는 트라이그램이며 4이상일 때는 gram 앞에 숫자를 붙여서 사용한다. 레벤슈타인 거리는 두 개의 문자열이 어느 정도 다른지를 나타내는 것이다. 편집거리를 계산하여 철자 오류 수정, 비슷한 어구 검색 등에 사용된다[4].

3. 제안하는 시스템

제안하는 클라우드 환경에서 자연어처리 기법을 활용한 취약점 분석 시스템은 도커 컨테이너 이미지를 수집하고, 수집된 도커파일과 공식적으로 등록되어있는 도커파일에 n-gram 문서 유사도 기법을 적용하고 유사도가 낮은 문서에 대해서는 레벤슈타인 거리 기법을 통해 다른 부분을 분석하고 제공한다. 다음 [Fig. 2]은 제안하는 시스템의 구조를 나타낸다.



[Fig. 2] Proposed system structure

4. 결론 및 향후 연구

본 논문에서는 클라우드 환경에서 자연어 처리 기법을 활용한 취약점 진단 시스템 설계를 제안하였다. 제안하는 클라우드 환경에서 자연어 처리 기법을 활용한 취약점 진단 시스템은 자연어 처리 기법을 활용한 취약점 분석 모듈을 통해 클라우드의 도커파일 데이터에서 취약점과 연관되는 값들을 분석하는 시스템이다. 향후 본 논문에서 제안하는 시스

템을 활용한 구축이 필요하다.

ACKNOWLEDGMENT

“본 연구는 과학기술정보통신부 및 정보통신기획평가원의대학ICT연구센터지원사업의 연구결과로수행되었음”(IITP-2020-2020-0-01602)

참고문헌

[1] 엄상현, 김직수. "도커 컨테이너 가상화 기반 클라우드 환경에서 효율적인 스토리지 자원관리 방법." 한국정보과학회 학술발표논문집 . (2020): 1247-1249.

[2] 문주현, 김상훈, 신용태. "Apache Spark를 활용한 쿠버네티스 클라우드 취약점 진단 시스템 설계." 한국컴퓨터정보학회 학술발표논문집 28.2 (2020): 543-544.

[3] 이모세, 강민수, 김인호, 김재헌. "실시간 분석을 위한 도커 컨테이너 기반의 딥러닝 모델 관리 시스템 설계 및 성능 비교." 한국통신학회논문지 46.2 (2021): 390-400.

[4] 김동현, 김강석. "N-gram을 활용한 DGA-DNS 유사도 분석 및 APT 공격 탐지." 정보보호학회논문지 28.5 (2018): 1141-1151.