Knowledge Graph 기반 RAG를 활용한 LLM 취약점 탐지 방법론

황시준* 이연준**
*한양대학교 정보보호학과(대학원생), **한양대학교 컴퓨터공학과(교수)
*ghkdtlwns987@hanyang.ac.kr, **yeonjoonlee@hanyang.ac.kr

Knowledge Graph based RAG for Vulnerability Detection with Large Language Models

Sijune Hwang*, Yeonjoon Lee**
**Hanyang University, Department of Computer and Information Security (Master Student)
***Hanyang University, Computer Science and Engineering (Associate Professor)

요 약

대규모 언어 모델(LLM)은 소프트웨어 보안 분야에서 새로운 가능성을 열며, 특히 소스코드 내 취약점을 자동으로 탐지하는 데 활발히 적용되고 있다. 그러나 LLM 기반 탐지 기법은 환각(Hallucination)으로 인한 오탐(False Positive) 문제를 안고 있어 실제 환경에 적용하기 어렵다는 한계가 존재한다. 이러한 문제를 완화하기 위해 Retrieval-Augmented Generation(RAG) 기법이 도입되었지만, Common Weakness Enumeration(CWE) 취약점 간의 관계를 충분히 반영하지 못하는 한계를 가진다. 본 연구는 이러한 한계를 극복하기 위해 CWE 관계를 Knowledge Graph(KG) 형태로 정의하고, 이를 RAG 와 결합한 새로운 탐지 방법론을 제안한다. 이는 CWE 간의 연계 시나리오를 효과적으로 식별하며, 기존 방법론 대비 더 정밀하고 설명 가능한 탐지 결과를 제공하였다. 본 연구는 단일 취약점 탐지를 넘어 복합적 취약점 시나리오 분석을 가능하게 한다는 점에서 차별성을 가진다.

Ⅰ. 서 론

LLM 은 소프트웨어 보안 연구 분야 전반에 채택되고 있으며, 소스 코드 내 취약점 자동 탐지에 유망한 성과를 보이고 있다[5][6][9]. 그러나 실제 적용 단계에서 환각[2] 문제로 인한 오탐을 유발하여 신뢰도가 저하되고, 이는 보안 도메인의 특성상 도입을 어렵게 만드는 핵심 문제이다. 환각은 LLM 이 훈련 데이터나 프롬프트 맥락에 없는 사실을 만들어내는 현상으로, 탐지 근거의 부재와 일관성 없는 설명을 동반하여 운영 상 리스크를 키운다. 이 문제를 완화하기 위한 접근으로 RAG[1]가 주목받고 있다. RAG 는 외부 지식을 참조해 응답을 보강함으로써 환각을 줄인다. 하지만 취약점 탐지라는 특수 과제에서는 텍스트 위주 검색만으로는 한계가 분명하다. 실제 보안 사고는 단일 취약점의 유무보다, 취약점 간의 인과관계로 전개되는 시나리오가 중요하다. 예를 들어 입력 검증 부재(CWE-20)가 선행되면, 이후 신뢰되지 않은 데이터 역직렬화(CWE-502)로 확장/연계될 위험이 커진다. 그러나 일반적인 RAG 파이프라인은 이런 구조적 관계를 명시적으로 반영하지 못해, “무엇이 왜 위험한가”를 연쇄적으로 설명하는 능력이 제한된다. 본 논문은 이러한 간극을 메우기 위해, CWE 취약점과 그 상호 관계를 그래프로 구조화한 Knowledge Graph(KG)[3][8]를 정의하고 이를 RAG 와 결합하는 KG 기반 RAG 취약점 탐지 방법론을 제안한다.

Ⅱ. 본론

RAG 는 LLM 의 환각을 완화하지만, 텍스트 중심 유사도에 의존하여 취약점 간 구조적 관계를 활용하지 못하며, 선행 연구로 Vul-RAG [4]가 있다. 이는 코드 의미 정보를 기반으로 검색을 수행해 취약점 탐지 성능을 개선했으나, CWE-20 이 CWE-502 로 이어질 수 있는 취약점 간의 의존성을 명시적으로 반영하지 못한다. 따라서 기존 방법론은 개별 취약점 탐지에 그치며, 실제 보안 시나리오를 설명하기에는 부족하다. 본 연구는 이러한 문제를 해결하기 위해 KG 기반 RAG 탐지 기법을 제안한다. KG 는 개체(Entity)와 관계(Relation)를 그래프 형태로 구조화 하는 지식 표현 방식으로, 본 연구에서는 CWE 취약점과 그 상호 관계(ChildOf, PeerOf, CanPrecede 등)를 모델링하여 구축한다[3]. 이를 통해 LLM 은 단순히 개별 CWE 존재 여부만 판별하는 것이 아니라, 어떤 취약점이 다른 취약점으로 확장되거나 선행될 수 있는지를 추론할 수 있다.

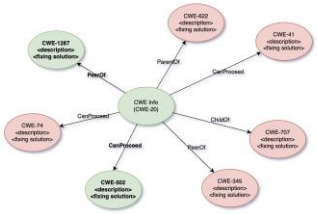

그림 1. CWE Knowledge Graph

제안 방법은 다음 세 단계로 구성된다.

1. Knowledge Graph 구축: CWE 데이터와 관련 문헌을 기반으로 취약점 유형과 그 관계를 구조적으로 정의하여 KG 를 생성한다.

2. 질의 확장: LLM 이 탐지를 수행할 때, KG 에서 제공하는 취약점 간 관계 정보를 질의에

반영한다. 예를 들어, CWE-20 이 탐지된 경우 KG 를 참조하여 CWE-1287, CWE-502 간의 연계 가능성을 함께 질의한다.

3. RAG 기반 탐지: KG 가 제공하는 구조적 지식을 외부 지식으로 활용하여, LLM 이 단일 취약점 뿐 아니라 취약점 시나리오 전체를 추론할 수 있도록 한다. 이는 단일 탐지 정확도를 높일 뿐 아니라, 복합 취약점에 대한 맥락적 설명을 제공한다.

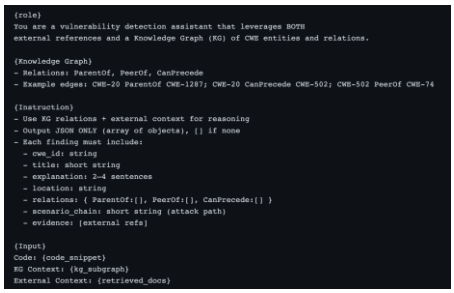

그림 2. KG + RAG prompt

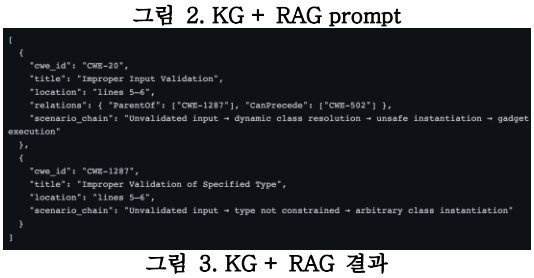

그림 3. KG + RAG 결과

제시한 그림 2 와 그림 3 은 제안한 KG + RAG 기반 탐지 과정의 실제 예시를 보여주며 CWE-20 과 CWE-502 이 연계된 사례로, KG 에 정의된 관계 정보를 활용하여 두 CWE 간의 연관성을 명확히 식별하였다. 이를 통해 단일 취약점 탐지를 넘어 구체적인 취약점 시나리오를 도출할 수 있음을 확인했으며, 취약점 간 관계를 중심으로 한 탐지와 시나리오 식별이 가능함을 시사한다. 또한, KG 는 새로운 CWE 관계나 보안 패턴이 발견될 경우 손쉽게 확장 가능하므로, 지속적으로 진화하는 취약점 탐지 프레임워크를 제공할 수 있으리라 기대한다.

Ⅲ. 결론

본 논문은 LLM 기반 취약점 탐지에서 발생하는 오탐과 환각 문제 [2], 그리고 기존 RAG 접근법 [1][4]의 한계를 지적하고, CWE 를 중심으로 한 취약점 관계 KG 를 구축하여 이를 RAG 와 결합하는 새로운 탐지 방법론을 제안하였다. 제안된 방법은 단일 CWE 탐지를 넘어, 취약점 간 관계와 시나리오를 식별함으로써 탐지 결과의 정확성과 설명 가능성을 동시에 향상시킬 수 있을 거라 기대한다. 향후 연구에서는 KG 자동 구축과 대규모 취약점 데이터셋에 대한 실증적 평가를 수행 함으로써 본 방법론의 실용성과 범용성을 강화할 계획이다.

ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-대학 ICT 연구센터(ITRC)의 지원을 받아 수행된 연구임(IITP-2025-RS-2024-00438056)

참 고 문 헌

[1] Lewis, Patrick, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." *Advances in neural information processing systems* 33 (2020): 9459-9474.

[2] Ji, Ziwei, et al. "Survey of hallucination in natural language generation." *ACM computing surveys* 55.12 (2023): 1-38.

[3] Chen, Xiaojun, Shengbin Jia, and Yang Xiang. "A review: Knowledge reasoning over knowledge graph." *Expert systems with applications* 141 (2020): 112948.

[4] Du, Xueying, et al. "Vul-rag: Enhancing llm-based vulnerability detection via knowledge-level rag." *arXiv preprint arXiv:2406.11147* (2024).

[5] Zhou, Xin, et al. "Large language model for vulnerability detection and repair: Literature review and the road ahead." *ACM Transactions on Software Engineering and Methodology* 34.5 (2025): 1-31.

[6] Zhou, Xin, Ting Zhang, and David Lo. "Large language model for vulnerability detection: Emerging results and future directions." *Proceedings of the 2024 ACM/IEEE 44th International Conference on Software Engineering: New Ideas and Emerging Results*. 2024.

[7] Zhou, Xin, et al. "Comparison of static application security testing tools and large language models for repo-level vulnerability detection." *arXiv preprint arXiv:2407.16235* (2024).

[8] Wen, Yilin, Zifeng Wang, and Jimeng Sun. "Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models." *arXiv preprint arXiv:2308.09729* (2023).

[9] Tamberg, Karl, and Hayretdin Bahsi. "Harnessing large language models for software vulnerability detection: A comprehensive benchmarking study." *IEEE Access* (2025).