International Journal of Information Security and Privacy
Volume 18 • Issue 1

VCGERG:
Vulnerability Classification With Graph Embedding Algorithm on Vulnerability Report Graphs

Yashu Liu, Beijing University of Civil Engineering and Architecture, China
https://orcid.org/0000-0002-7208-5360

Xiaoyi Zhao, Beijing University of Civil Engineering and Architecture, China*

Xiaohua Qiu, Beijing University of Civil Engineering and Architecture, China

Han-Bing Yan, National Computer Network Emergency Response Technical Team, China

ABSTRACT

Vulnerability can lead to data loss, privacy leakage and financial loss. Accurate detection and identification of vulnerabilities is essential to prevent information leakage and APT attacks. This paper explores the possibility of digging the valuable information in vulnerability reports deeply. We propose a new model, VCGERG, which products a graph using key information from vulnerability reports and embeds the graph into the vector space using a keywords-LINE graph embedding algorithm based on the attention of neighboring nodes. VCGERG model uses the OVR random forest algorithm to classify vulnerabilities. Our model can get the complicated local and global information of the graph in large-scale dataset and achieve better results. In order to verify the effectiveness of our model, it is evaluated on many experiments. Compared with other models, our method has a higher accuracy rate of 0.975.

KEYWORDS

Graph Embedding, Graph Structure, Vulnerability Classification, Vulnerability Reports

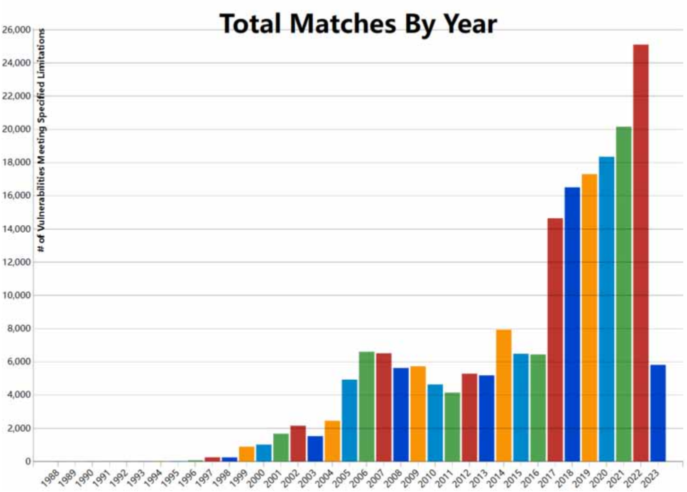
Vulnerability is a flaw in the specific implementation of hardware, software, protocols, or system security policies. It is used by attackers to access or destroy the system without authorization. The National Vulnerability Database (NVD) (National Institute of Standards and Technology [NIST], 2023a) provides data on the subject as shown in Figure 1. The figure shows that the total number of vulnerabilities reached 25,102 in 2022, a significant increase of 24.5% compared with the previous year. Thus, the escalating severity of cybersecurity issues has prompted researchers to focus on identifying and classifying these vulnerabilities. However, the vast number and heterogeneity of vulnerability information pose a significant challenge to be able to detect and classify vulnerabilities quickly.

Certain institutions have now organized identified vulnerabilities. Common Weakness Enumeration (CWE) (2023) provides an extensive list of prevalent security weaknesses, where each

DOI: 10.4018/IJISP.342596

*Corresponding Author

Figure 1. Amount of Vulnerability Data Released by NVD



one corresponds to a different class of software bugs or defects. The latest CWE list, version 4.14, lists 1,362 different CWE-IDs and provides a comprehensive framework for identifying and classifying common security weaknesses. Researchers can facilitate systematic examination and analysis of the vulnerabilities using this framework. Attackers exploit vulnerabilities taking advantage of bugs or flaws in software systems. For example, microweber security vulnerability (CVE-2022-2353) (NIST, 2023b) appeared in July 2022. It is classified as a "cross-site request forgery" (CSRF) vulnerability (CWE-352), which allows an attacker to steal cross-site request forged tokens, access content from the same site, and redirect. This vulnerability poses risks such as data leakage and user privacy losses. Another notable example is the Apple iOS and macOS out-of-bounds write vulnerability (CVE-2022-32893) (NIST, 2023c), which appeared in August 2022. It is classified as an "out-of-bounds write" vulnerability (CWE-787).

In addition to CWE, vulnerability reports released by the NVD and security vendors have significant value in understanding and discovering vulnerability characteristics. The reports include summary information, the operating system and software type, threatening behavior, MD5, and other text information. The reports are unstructured and consist of text and tables. We can extract and reorganize the information from a vulnerability report. We can then describe vulnerability features from a more diverse perspective and construct a vulnerability feature library, which can be used to detect vulnerabilities automatically. In security and confidentiality applications, timely deep learning analysis and classification of data and information are of great significance (Aljarf et al., 2023; Altalhi & Gutub, 2021; Sufi et al., 2023). Using machine learning and deep learning methods further improves the accuracy of the classification results compared with traditional methods (Aljarf et al., 2023; Gutub et al., 2023; Hemalatha et al., 2023; Roy et al., 2023). This paper is dedicated to obtaining and analyzing the information in vulnerability reports and classifying the vulnerabilities.

We propose a graph structure to represent heterogeneous information in vulnerability reports, incorporate the LINE algorithm based on the attention mechanism of neighboring nodes to characterize vulnerabilities, and use the one VS rest (OVR) random forest model to identify the vulnerabilities.

Our three main contributions with this research are as follows:

1. Vulnerability report graph representation method. We propose an unstructured and heterogeneous information representation method using a graph. It breaks the limitation of the traditional approach