Global Semantic-based Code Defect Detection

Abstract

Software technology is becoming more and more closely related to all aspects of social life, and The software development inevitably produces various loopholes, and the identification and positioning of loopholes is very labor-intensive and resource-intensive. How to quickly identify and locate vulnerabilities to improve the stability and security of software operation has become an increasingly serious problem. With the development of deep learning technology, there are some methods that can quickly and automatically identify software vulnerabilities, such as vulnerability identification methods based on code abstract syntax tree (hereinafter referred to as AST) or program data flow graph (hereinafter referred to as PDG), from AST or PDG However, most of the code organization structure will disappear after abstracting through AST, which makes it difficult to capture the semantics of interdependence between code elements, which is not conducive to the identification of vulnerabilities. This paper proposes a code vulnerability feature extraction method based on lexical analysis to exploit wider semantic dependencies for vulnerability identification.

软件技术与社会生活的方方面面越来越紧密，软件开发又不可避免的产生各种漏洞，而漏洞识别与定位十分耗费人力物力。如何快速识别并定位漏洞以提升软件运行的稳定性与安全性，成为越来越严峻的问题。随着深度学习技术的发展，出现了一些可快速自动识别软件漏洞的方法，譬如基于代码抽象语法树（以下简称AST）或程序数据流图(以下简称PDG)的漏洞识别方法，从AST或PDG中提取代码特征进行漏洞模式识别，但大部分代码组织结构譬如通过AST进行抽象后会消失，导致代码元素间相互依赖的语义难以捕捉，不利于漏洞的识别。本文提出一种基于词法分析的代码漏洞特征提取方法，以利用更广的语义依赖进行漏洞识别。