論 文 内 容 の 要 旨

氏 名 （ Pizzolotto Davide ）

論文題名	Scalable Clone Detection on Low-Level Codebases (低水準言語コードに対する大規模なコードクローン検出)

論文内容の要旨

In recent years, the adoption of open source software and its inclusion into closed source projects greatly increased.
This does not only include open source libraries, but also code snippets copied and pasted from Stack Overflow.
Aside from the potential license violation, this code has been proven often outdated or containing vulnerabilities that have long been patched in the original code.

While most clone detection tools are focused on detecting copied code in source code, little work exists that can be applied to binary code or lower-than-source code.
This may be a problem in case the original source code is not available or in case of huge codebases to analyze.
In fact, existing binary clone detectors are usually limited to pairwise analysis.

In this dissertation, we discuss the detection of cloned snippets in low-level code. The dissertation is divided into two main parts: a first part highlighting techniques to improve clone detection in source code by transforming code into its low-level counterpart, and a second part focusing on scalable binary clone detection.

The first part provides two different approaches to improve clone detection by means of low-level code.
In the first approach we take programs written in the Rust language, a particular language with a compilation pipeline composed by multiple steps, and we apply regular clone detection after performing some compilation transformations.
While this approach is specific to a particular language, the second one is more generic and consist of implementing transformations similar to the one of a compiler in order o provide code normalization and improve clone detection.

The second part of this dissertation is instead based exclusively on binary code detection and is subdivided into two sub-parts. First, we discuss a novel approach for detecting clones in a binary file in a scalable way.
This novel approach uses methods commonly found in decompilation and builds a representation of functions that can be compared in linear time.
We show, however, that this method is highly sensitive of optimization options and, for this reason, we close this second part of the dissertation with a study on optimization flags detection using learning approaches.

Ultimately, we provide different methods for detecting clones in low-level languages using scalable approaches: in particular, in binary code clone detection we show that it is possible to reach the same precision of existing learning approaches while keeping the speed of a traditional one that is an order of magnitude faster.

様式７

論文審査の結果の要旨及び担当者

氏　名　（　Davide Pizzolotto　）			
論文審査担当者		（職）	氏　名
	主 査	教授	肥後芳樹
	副 査	教授	楠本真二
	副 査	教授	伊野文彦

論文審査の結果の要旨

近年，商用プロジェクト等におけるクローズドソースソフトウェアにおいてオープンソースソフトウェアの採用が進んでいる．このような再利用は，オープンソースライブラリだけでなく，Stack OverflowなどのＱＡサイトからのコピー&ペーストされたコード片も含んでいる．このような重複コード（以降，クローン）は，ライセンス違反の可能性もさることながら，古いコードであったり，オリジナルのコードでは修正済みの脆弱性を含んでいたりすることが先行研究により示されている．ほとんどのクローン検出ツールは，ソースコード内のコピーされたコードを検出することに重点をおいており，バイナリコードやローレベルコードに適用できるツールはほとんど存在しない．これは，オリジナルのソースコードが入手できない場合や，解析対象のソースコードが巨大な場合に問題となる．実際に，既存のバイナリクローン検出器はペアワイズ解析に限定されている．

本論文では，ローレベルコードにおけるクローンの検出について議論している．本論文は2つの主要な部分に分けられる．第1部はソースコードをローレベルコードに変換することによってソースコード中のクローン検出を改善する技術に焦点を当て，第2部はスケーラブルなバイナリクローン検出に焦点を当てている．

第1部ではローレベルコードを用いることによってクローンの検出精度を向上させる2つの異なるアプローチを提案している．最初のアプローチでは，複数のステップからなるコンパイルパイプラインを持つプログラミング言語であるRustで書かれたプログラムを取り上げ，いくつかのコンパイル変換を行った後に，通常のクローン検出を適用している．このアプローチが特定のプログラミング言語に特化しているのに対して，第二のアプローチはより汎用的であり，コードの正規化とクローン検出の改善を目的として，コンパイラに類似した変換を実装することで構成されている．

本論文の第2部は，バイナリコード検出に特化しており，2つのサブパートに分かれている．まず，バイナリファイル中のクローンをスケーラブルに検出するための新しいアプローチについて述べている．この新しいアプローチは，デコンパイルで一般的に利用される手法に倣い，線形時間で比較可能な関数の表現を構築することができる．しかし，この方法は最適化オプションに非常に敏感であるため，学習アプローチを用いた最適化オプションの検出に関する手法についても本論文は含んでいる．提案したバイナリコードクローン検出では，従来の学習型アプローチの速度を維持したまま，既存の学習型アプローチと同じ精度に到達することが可能であり，その速度は1桁以上速いことが実験により示された．

本論文の主な貢献はバイナリの差分に対する新しいアプローチであり，既存手法と同程度の精度を保ちつつ，自動的かつ高いスケーラビリティを実現していることである．既存手法は主に手動での解析に基づいているか，一組のバイナリに対して長い実行時間を必要としている．一方，本論文のアプローチは既存手法に比べて最大 1000 倍高速であり，数百メガバイトまでのバイナリを対象とすることができ，完全に自動で検出を行うことができ，既存手法と同程度の精度が得られることが実験により示されている．これらの結果から，LLVM や GCC のような巨大な対象においても，既存の手法では数ヶ月から数年を要するバイナリの差分を本論文の手法を用いることにより短時間で得られるようになった．これにより，ソフトウェアの進化や脆弱性・マルウェアの伝播において新たなフロンティアを開拓することができる．

以上のことから，博士（情報科学）の学位論文として価値のあるものと認める．