



Title	Fast and Precise Token-Based Code Clone Detection
Author(s)	村上, 寛明
Citation	大阪大学, 2016, 博士論文
Version Type	VoR
URL	https://doi.org/10.18910/55840
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

<https://ir.library.osaka-u.ac.jp/>

The University of Osaka

論文内容の要旨

氏 名 (村上 寛明)

論文題名

Fast and Precise Token-Based Code Clone Detection
(高速かつ高精度な字句単位のコードクローン検出)

論文内容の要旨

近年、ソフトウェア工学の分野においてコードクローン（以降、クローン）が注目を集めている。クローンとはソースコード中に存在する同一または類似するコード片を指し、主にコピーアンドペーストによって生成される。クローンはソフトウェア保守を困難にする要因であると言われている。その理由は、開発者があるコード片に変更を加える場合、開発者はそのコード片のクローンにも同様の修正が必要であるかを検討しなければならないためである。一方でコピーアンドペーストによる既存コードの再利用はいくつかの利点をもつ。その一つとして、既に存在する機能と類似した機能を素早く実装できることが挙げられる。

クローンをソースコード中から自動的に検出するために、これまでに様々なクローン検出ツールが開発されている。現在のところクローンの厳密な定義は存在しないため、各開発者が独自のクローンの定義を定め、その定義に基づいてクローン検出ツールを開発しているという現状である。そのような背景からクローン検出ツールの精度やパフォーマンスを比較する研究も行われている。その研究結果から、私は既存のクローン検出ツールには二つの課題点があると考えた。一つ目は検出時間が長いことであり、二つ目は検出精度が十分でないことである。検出時間が長いという課題点はプログラム依存グラフを用いた検出手法に見られるものである。プログラム依存グラフを用いた検出では、まずソースコードをプログラム依存グラフに変換し、その後、同型部分グラフをクローンとみなす。この検出手法は同型部分グラフを見つけるのに多くの時間を要する。次に、検出精度が十分でないことの原因として以下の二つが挙げられる。一つ目は、モジュール単位の検出手法に見られるものである。モジュール単位の検出手法では、類似したモジュール（ブロックやメソッド等）をクローンとみなす。そのためモジュールの内部で部分的に重複している場合、それを検出することはできない。二つ目は、似た命令が繰り返し記述された箇所起因するものである。似た命令が繰り返し記述された箇所とは、switch文に含まれる連続したcaseエントリや、連続した類似関数呼び出しを指す。ここで、クローン検出手法の中に字句単位の検出手法というものが存在する。字句単位の検出手法は、ソースコード中において連続して一致した二つの字句列をクローンとみなす。字句単位の検出手法を用いて似た命令が繰り返し記述された箇所からクローン検出を行うと、開発者にとって興味のないクローンが多く検出される。興味のないクローンの大量検出は検出精度を下げる原因である。

本研究では、これらの課題点を改善する二つのクローン検出手法を提案した。一つ目の手法は、似た命令が繰り返し記述された箇所を折りたたんだ後に、字句単位の検出手法を用いてクローンを検出する。似た命令が繰り返し記述された箇所の折りたたみは、開発者にとって興味のないクローンが検出されるのを抑制する。二つ目の手法は、Smith-Watermanアルゴリズムを用いてクローンを検出する。この検出手法はグラフを用いていないため高速な検出が可能となり、またモジュールより細かい粒度での検出を行うためモジュールの内部で部分的に重複している箇所もクローンとして検出可能である。Bellonらの実験に基づいて二つの提案手法を評価した結果、提案手法は既存手法の課題点を改善していることを確認した。

また本研究では、提案した検出手法を用いて検出されたクローンを可視化する手法を提案した。前述したように、開発者があるコード片に変更を加える場合、開発者はそのコード片のクローンにも同様の修正が必要であるかを検討しなければならない。その状況において、開発者はソースコード群および変更を加える対象となるコード片を入力し、ソースコード群に含まれるそのコード片のクローンを出力するツールをしばしば利用する。しかし検出された全てのクローンを閲覧する場合、ツールの使用者は複数のソースファイルを開かなければならなかったり、スクロールバーを操作する必要があったりする等の課題点が存在する。既存の可視化手法の課題点を改善するために、本研究ではCircle Packingを用いてクローンを可視化する手法を提案した。さらに前述したクローン検出手法とクローン可視化手法を組み合わせたツールを実装した。実装したツールの有効性を確かめるため、被験者実験を行った。被験者実験では、提案手法を実装したツールと既存ツールを対象として、クローン分析に要する時間およびユーザビリティを比較した。実験の結果、その双方について提案手法を実装したツールは既存ツールに比べて優れていることを確認した。

論文審査の結果の要旨及び担当者

氏 名 (村上寛明)			
	(職)	氏 名	
論文審査担当者	主 査	教授	楠本真二
	副 査	教授	萩原兼一
	副 査	教授	井上克郎
	副 査	准教授	肥後芳樹

論文審査の結果の要旨

本論文は、ソフトウェア開発や保守の支援技術として注目されているコードクローン分析の課題であるコードクローン検出時間の削減と検出精度の向上に取り組んだ結果をまとめている。コードクローンとはソースコード中に存在する同一、あるいは、類似するコード片のことを意味し、主にコード片のコピーアンドペーストにより生成される。近年、コードクローンはソフトウェア工学、特に、ソフトウェア保守の研究分野において注目を集めている研究テーマである。また、コードクローンが原因となり失敗した大規模ソフトウェア開発プロジェクトの事例も報告されており、実用面においても重要なテーマとなっている。

本論文では、複数あるコードクローン検出手法の中から、字句単位でコードクローンを検出する手法を対象としている。字句単位の検出手法は、ソースコード中において連続して一致した二つの字句列をコードクローンとみなすものであり、検出速度は他の検出手法よりも早い、開発者にとってあまり意味の無いコードクローンも検出してしまうという課題を持つものである。また、元のコード片をコピーアンドペーストした後で、若干の修正を加えたようなコードクローンの検出に適用できないという課題もある。

本論文では、これらの課題点を改善する二つのコードクローン検出手法を提案している。最初の提案手法は、似た命令が繰り返し記述された箇所を折りたたんだ後に、字句単位の検出手法を用いてコードクローンを検出する手法である。似た命令が繰り返し記述された箇所の折りたたみは、開発者にとって意味のないクローンが検出されるのを抑制する効果がある。次に、Smith-Watermanアルゴリズムを用いてコードクローンを検出する手法を提案している。この検出手法は、バイオインフォマティクス分野で遺伝子の塩基配列中から類似する部分配列のペアを検出する際によく用いられている。このアルゴリズムを用いることで、高速なコードクローン検出と部分的に重複している箇所をコードクローンとして検出可能とした。コードクローン研究分野で用いられているベンチマーキングデータを用いて二つの提案手法を評価した結果、提案手法は既存手法よりも検出速度、検出精度において向上が見られ、コードクローン検出における二つの課題の改善に寄与したことが確認された。

また本論文では、提案した検出手法を用いて検出されたコードクローンの可視化手法の提案とそれに基づく可視化ツールの開発も行っている。可視化ツールは、開発者が不具合の修正等を目的として、あるコード片に変更を加える場合、そのコード片のコードクローンにも同様の修正が必要であるかどうかを分析する際の支援をするものである。開発者が修正対象コード片と確認対象のソースコード群をツールに与えることで、修正対象のコードクローン情報が可視化されて提示される。可視化にあたっては、Circle Packingと呼ばれる可視化手法を応用している。コード修正の被験者実験の結果、修正時間、ツールの使用性について既存ツールと比較して有用であることが確認されている。不具合修正時に、同様の修正を残さず修正するという作業は実際の開発現場においてよく発生する作業であり、提案されている可視化ツールの有用性は高いものと判断できる。

以上のように、本論文の成果は、コードクローン分析を支援することにより、ソフトウェア開発・保守の効率化において、技術面、並びに、実用面における高い貢献があると考えられ、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。