

重複コードと非重複コードに加えられる修正の頻度の比較 -複数の検出ツールを用いた評価-

A Comparison of Modification Frequency between Duplicate Code and Non-Duplicate Code
-Evaluation with Multiple Detection Tools-

堀田 圭佑[†]
Keisuke Hotta

佐野 由希子[†]
Yukiko Sano

肥後 芳樹[†]
Yoshiki Higo

楠本 真二[†]
Shinji Kusumoto

大阪大学[†]
Osaka University

1 まえがき

一般的に、ソースコード中の同一もしくは類似するコード片(以下、重複コード)は、修正に要する作業量を増加させる恐れがあるといわれている。しかし、実際に重複コードが修正の作業量に与える影響を定量的に調査した研究は少ない[1][2]。

そこで、本稿では、重複コードに加えられる修正の頻度に注目する。重複コードに加えられる修正の頻度が高ければ、重複コードが修正に要する作業量を増加させている、と考えられる。

いくつかのオープンソースソフトウェアを対象に計測を行った結果、重複コードに加えられる修正の頻度が、非重複コードに加えられる修正の頻度よりも、やや低いという結果が得られた。

2 修正頻度計測手法の概要

2.1 計測方法

計測方法の概要を以下に示す。

1. 構成管理ツール(CVS, Subversionなど)の履歴情報から、リビジョンごとに変更されたファイルの情報を取得する。
2. 修正前後のファイルの差分から修正位置を特定する。
3. 修正前のファイルに重複コード検出ツールを使用し、重複コードとなっている部分を特定する。
4. 2, 3 の情報を元に、修正が重複コードに加えられたものか非重複コードに加えられたものかを特定する。

また、特定のツールに依存しない結果を得るために、CCFinder, CCFinderX, Simian, Scorpioの四種類の重複コード検出ツールを用いて計測を行う。

2.2 修正頻度の定義

ソースコードに変更が加えられたリビジョンを、計測対象リビジョンとする。

本稿では、以下の値を修正頻度と定義する。

$$\frac{\text{重複コードの修正箇所数}}{\text{対象リビジョン総数}} \times \frac{\text{総行数}}{\text{重複コード行数}}$$

なお、重複コード含有率(重複コード行数/総行数)の逆数をかけているのは、重複コード含有率の影響を取り除くためである。非重複コードの修正頻度は、重複コー

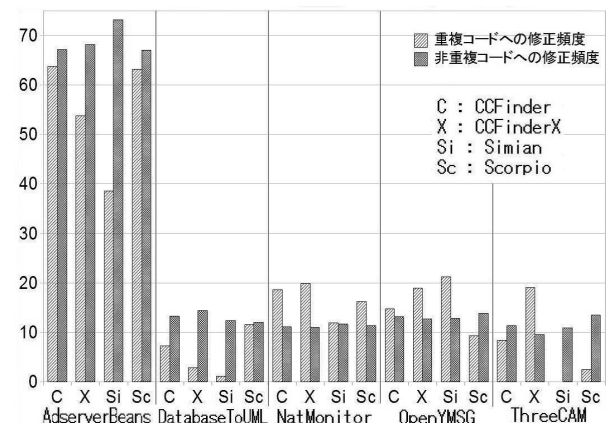


図1 計測結果

ドの修正箇所数および重複コード行数を、非重複コードの修正箇所数および非重複コードの行数に置き換えたものと定義する。

3 計測結果

五つのJavaプロジェクトを対象に計測を行った結果を図1に示す。

二十のデータの内、十二のデータにおいて、非重複コードの修正頻度が重複コードの修正頻度を上回るという結果が得られた。

4 まとめ

本稿では、重複コードが修正の作業量を増加させているのかを定量的に調べるため、複数の重複コード検出ツールを用いて重複コードと非重複コードの修正頻度を計測した。いくつかのオープンソースソフトウェアを対象に計測したところ、重複コードの修正頻度が非重複コードの修正頻度より低い、という結果を得た。

参考文献

- [1] 門田暁人, 佐藤慎一, 神谷年洋, 松本健一, "コードクローンに基づくレガシーソフトウェアの品質の分析", 情報処理学会論文誌, Vol.44, No.8, pp.2178-2188, Aug.2003
- [2] Angela Lozano, Michel Wermeliner, "Assessing the effect of clones on changeability", International Conference on Software Maintenance, pp.227-236, 2008