

Github における PullRequest を用いた言語別のプロジェクトの特徴分析

石田 凌 戸田 航史
福岡工業大学[†] 福岡工業大学[‡]

1. はじめに

近年オープンソースソフトウェア (OSS) の普及に伴い、多くの開発者が開発に携わるようになってきている。多くの OSS 開発、特に GitHub (<https://github.com>) 上で開発が行われているプロジェクトにおいては、プルリクエストを主体とするプル型ソフトウェア開発が実施されている。プル型開発では機能追加や不具合修正はプルリクエストという形で提供され、プロジェクトの進捗速度やクオリティに大きな影響を与える。本研究ではプルリクエストのマージの可否やその決定までの時間に言語が与える影響について比較、検証を行った。

2. 研究環境

本研究では解析対象として Gousios らが Github から収集したデータを用いた [1]。収集期間は 2012 年 1 月から 2013 年 8 月、20 万プロジェクト 190 万のプルリクエストが対象である。ここから以下の手順でデータの選定を行った。

1. 期間内でのプルリクエストが 200 未満のプロジェクトを除外する。
2. ソースコードにテストが含まれるプロジェクトだけを選定する
3. 開発言語が Ruby, Java, Python, Scala の 4 言語のプロジェクトだけを選定する。
4. 最低一度はコアメンバではない開発者からのプルリクエストがコミットされているプロジェクトを選定する。
5. ドキュメント作成, プログラミング言語実装のプロジェクトを除外する。

最終的に 291 プロジェクトの 16684 のプルリクエストを対象とすることにし、言語ごとの件数を表 1 に示す。

3. 実験概要

本研究ではプロジェクトを言語ごとに分け、言語ごとに表 2 の各項目の統計量の算出やその検定、RandomForest [2] による予測と重要度の

表 1. 言語別プロジェクト・プルリクエスト数

	プロジェクト数	プルリクエスト数
Scala	14	7,576
Python	99	59,970
Java	91	43,870
Ruby	87	55,468

表 2. 実験で用いたメトリクス

コミット数
変更行数
変更テスト行数
変更ファイル数
コメント数
レビュー参加者数
レビュー内の競合コメント数
他のプルリクエストへのリンク
SLOC
過去 3 ヶ月のコアメンバ数
過去 3 ヶ月の外部開発者参加率
過去 3 ヶ月の総ファイル変更数
テスト網羅度
開発者のプルリクエスト提出数
開発者のプルリクエストマージ率

算出を実施し比較を行った。

3.1 統計量の算出

言語ごとに Mergetime, Lifetime の統計量を算出した。ここで Mergetime とはプルリクエストがマージされるまでの時間、Lifetime はプルリクエストがマージ、または拒否されるまでの時間である。

3.2 RandomForest の実施

Mergetime, Lifetime を目的変数、表 2 の各項目を説明変数として RandomForest による予測を行い、重要度 (importance) を算出した。また、プルリクエストの受理に影響を与える変数を調査するために Merge_decision を目的変数とした実験も行った。Merge_decision とは提出されたプルリクエストの可否の判断である。

Feature analysis of project of each language by pull request in GitHub

[†] Ryo Ishida, Fukuoka Institute of Technology

[‡] Koji Toda, Fukuoka Institute of Technology

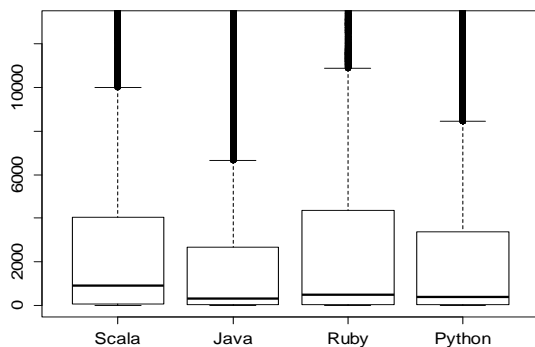


図 1. 言語別 Mergetime

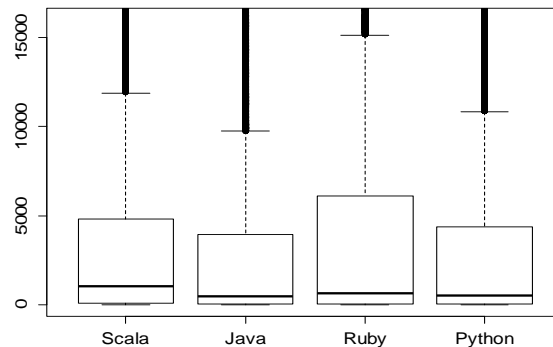


図 2. 言語別 Lifetime

4. 実験結果

4.1 統計量の比較

Mergetime, Lifetime に関する統計量の比較の箱ひげ図を図 1, 図 2 に示す. このグラフにおいて横軸は各言語を, 縦軸は時間 (分) を表している. どちらにおいても Ruby の時間が比較的長く, Java, Python は短いことが分かった. また, 有意差検定の結果 Mergetime の Scala-Java 間を除いた全ての組み合わせにおいて, 有意水準 99% で差が確認できた.

4.2 RandomForest における重要度

RandomForest の結果 Merge_decision を目的変数とした場合には各言語とも 過去 3 カ月間の総ファイル変更数, SLOC, テスト網羅度の順で上位を占めていたが, それ以下の説明変数との間にそれほど大きな差は見られなかった.

Mergetime においてはどの言語でも過去 3 カ月間の総ファイル変更数, SLOC, プルリクエスト提出者のプルリクエストマージ率, テスト網羅度が上位であった. ただし Ruby においてはプルリクエスト提出者のプルリクエストマージ率が他の 3 言語と比べて低く, Scala においては過去 3 カ月間の総ファイル変更数の重要度が他の 3 言語と比較して低いという結果が表れた.

Lifetime においては 4 言語で共通して SLOC, テスト網羅度, 過去 3 ヶ月の外部開発者参加率の順で重要度が高かった. その他の要因にも大きな違いは無かったが, Mergetime で上位であった過去 3 ヶ月間の総ファイル変更数は下位であった. この事から過去の総ファイル変更数はマージまでの時間には影響を与えるがマージの可否決定にかかる時間にはそれほど影響を与えないことが分かった.

5. おわりに

本研究ではプル型 OSS 開発においてプルリクエストの言語ごとの特徴を明らかにする事を目的として分析を行った.

分析の結果, Mergetime, Lifetime ともに言語間で有意差が見られたがその要因の重要度には大きな違いが見られなかった. ただし Ruby ではプルリクエスト提出者のマージ割合の重要度が低く, Scala では過去の総ファイル変更数の重要度が低い結果となった. Ruby については, そもそも Ruby の開発者はプルリクエストがマージされる割合が他言語に比べて低い傾向にあり, Scala については過去の総ファイル変更数がほかの 3 言語と比較して多く, これらが重要度に影響を与えた可能性がある. また, Merge_decision においては言語間で差異は見られなかった.

これらの結果から GitHub 上のプル型開発においては Ruby でマージの可否判断が長く Java や Python で短い傾向にあり, 過去 3 ヶ月間のファイルへの総変更数や SLOC がプロジェクトの Mergetime, Lifetime に大きな影響を与えることが分かった.

今回の研究は GitHub 上の 4 言語のみを対象としており今後は更に対象を広げて検証する必要があると考えている.

参考文献

- 1) G. Gousios, M. Pinzger, and A. van Deursen. : An exploratory study of the pull-based software development model. Proc. 36th International Conference on Software Engineering (ICSE 2014).
- 2) Leo. Breiman. : Random Forests. pp. 5-32 (2001)