2M-7

GitHub における Pull Request のマージに影響を与える要因の分析

梅本　祥平，神谷　年洋 †
公立はこだて未来大学 †

1　はじめに

分散型バージョン管理システムの Git を使用したコード共有サイトである GitHub(https://github.com/) では多数のオープンソースソフトウェア (OSS) が開発されている．GitHub における OSS の開発では，新機能の追加やバグの修正といったコードの変更を含むパッチを送信するための機能である Pull Request が使用される．Pull Request は送信された後に，取り込み (マージ) に関して開発者間で議論が行われ，マージの成否が決定される．このため，Pull Request には，マージされるものとマージされないものが存在する (図 1).

マージされない Pull Request は，開発者の作業の無駄につながる．これは開発者のモチベーションを低下させる可能性があり，特性の OSS の品質の低下や開発の中断を引き起こす可能性がある．

本研究では，Pull Request のマージに影響を与える要因を，マージに関する議論に着目して分析する．また，分析の結果から，マージされない Pull Request を減少させるための方法を検討する．

2　関連研究

これまでの研究から明らかとなった，Pull Request のマージに影響を与える要因と，本研究での分析の対象となるデータセットについて述べる．

2.1　Pull Request に含まれるテストコード

Pham ら [3] は，GitHub などのコード共有サイトにおけるテスト文化について調査を行った．開発者へのインタビューの結果から，Pull Request に含まれるテストコードが，Pull Request のマージに影響を与える要因であることを述べている．

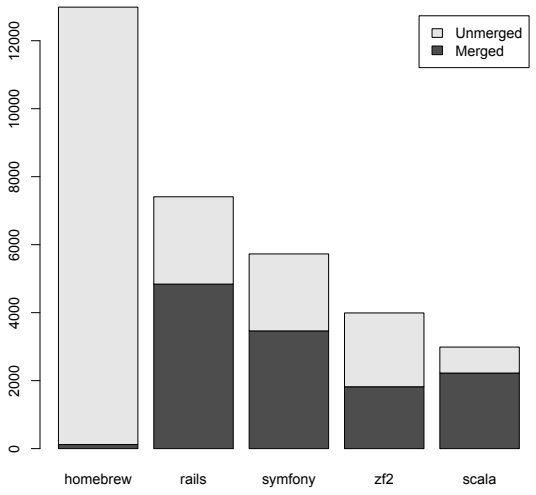

図 1　各プロジェクトにおけるマージされた Pull Request とマージされなかった Pull Request の割合

2.2　Pull Request を使用した開発モデル

Gousios ら [2] は，Pull Request を使用した開発モデルに関する探査的な研究を行った．Pull Request のマージに影響を与える要因として，Pull Request によって影響を受ける領域と Pull Request を送信するプロジェクトの規模，Pull Request によって変更されるファイル数を挙げている．

2.3　分析の対象とするデータセット

Gousios ら [1] は，GitHub API が提供する GitHub の様々な情報を収集し，研究に向けたデータセットを提供するプロジェクトである GHTorrent を展開している．

Merged Pull Requests and Unmerged Pull Requests
†Shouhei UMEMOTO, Toshihiro KAMIYA, Future University Hakodate

本研究では，分析の対象として，GHTorrent から提供されるデータセットを使用する．

3　提案方式

Pull Request のマージに影響を与える要因を分析する方法を述べる．まず，Pull Request が保持する状態について述べる．次に，Pull Request のマージに影響を与える要因が含まれる可能性がある議論を抽出する方法を述べる．

3.1　Pull Request の 2 つの状態

Pull Request は，Opened と Closed のどちらかの状態にある．Pull Request は，作成された時点で，まず Opened の状態に遷移する．Pull Request が Opened の状態である間に，マージの成否を決定するための議論が行われる．マージの成否を決定するための議論が終了し，マージの成否が決定されると，Pull Request は Closed の状態に遷移し，マージされるものかマージされないものとなる．一度，Closed の状態に遷移した Pull Request は，再度，Opened の状態に遷移することができる．Pull Request に関する議論が終了すると，再度，Closed の状態に遷移する．

3.2　Closed に遷移する直前と直後の議論の抽出

Pull Request が Closed に遷移する直前と直後の議論には，Pull Request のマージの成否の理由や結果が書かれていることが多い (図 2)．そこで，Pull Request が Closed に遷移する直前と直後の議論を抽出する．抽出した議論を収集し，単語の頻度分析を行うことで，マージに影響を与える要因を明らかにする．

4　評価と考察

分析から明らかになった要因に基づいて，Pull Request をマージされると予測されるものと，マージされないと予測されるものに分類する．予測の精度から，分析から明らかになった要因の有効性を評価する．

ただし，手動でマージされたものは，データセットではマージされていないものと判断されている．また，部分的にマージされたものも同様に，データセットではマージされていないものと判断されている．

図 2　Closed の直前にマージの成否の理由が含まれる議論

5　まとめ

本研究では，GitHub における Pull Requet のマージに影響を与える要因を，マージの成否を決定するために行われた議論を分析する手法を提案した．特に，Pull Request が Closed に遷移する直前と直後の議論を抽出することによって，マージに影響を与える要因を明らかにすることを述べた．

今後は，より詳細な分析を行うことで，マージされない Pull Request を減少させるための方法を確立する．また，データセットにおけるマージの成否と実際のマージの成否との相違を検出する手法を検討する．

参考文献

[1] G. Gousios. The ghtorrent dataset and tool suite. In *Proceedings of the 10th Working Conference on Mining Software Repositories*, MSR'13, pages 233–236, 2013.

[2] G. Gousios, M. Pinzger, and A. van Deursen. An exploration of the pull-based software development model. sep 2013. Submitted to the International Conference on Software Engineering 2014.

[3] R. Pham, L. Singer, O. Liskin, F. F. Filho, and K. Schneider. Creating a Shared Understanding of Testing Culture on a Social Coding Site. In *Proceedings of the 35th International Conference on Software Engineering (ICSE 2013), pages 112 - 121, San Francisco, USA*, 2013.