

Commit Message Matters: Investigating Impact and Evolution of Commit Message Quality

Jiawei Li
Donald Bren School of ICS
University of California, Irvine
Irvine, USA
jiaw128@uci.edu

Iftekhar Ahmed
Donald Bren School of ICS
University of California, Irvine
Irvine, USA
iftekha@uci.edu

2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE),
pp806-817

発表者：松尾 和輝

背景 | コミットメッセージ

- コミット時に自然言語で記述されるメッセージ
 - プログラムの理解や保守を助けることができるソフトウェアの変更の文書化^[65]
 - 開発者が共同作業者に変更の背景を伝えるコミュニケーションチャンネルの1つ^[47]
- 質の高いコミットメッセージを記述することが重要
 - 変更内容と変更理由を含むコミットメッセージ^[65]
 - セマンティクスを考慮

課題

質の高いコミットメッセージがプロジェクトに与える影響が不明

[47] U. A. Mannan, I. Ahmed, C. Jensen, and A. Sarma, "On the relationship between design discussions and design quality: a case study of apache projects," in Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2020, pp. 543–555.

[65] Y. Tian, Y. Zhang, K.-J. Stol, L. Jiang, and H. Liu, "What makes a good commit message?" arXiv preprint arXiv:2202.02974, 2022.

目的とRQ

- 目的
 - 質の高いコミットメッセージに関する様々な調査
- RQ
 - RQ1：コミットメッセージで参照される課題報告/プルリクエストへのリンクは常にWhy情報を提供しているか？
 - RQ2：コミットメッセージの品質はソフトウェアの欠陥発生率にどのような影響を与えるのか？
 - RQ3：OSSプロジェクトにおけるコミットメッセージの品質は時間とともにどのように進化するのか？

RQ1 | リンク先のWhy情報の有無

- アプローチ
 - 先行研究のデータを利用
 - コミットメッセージにissue/prリンクを含む611件をデータセットとする（What/Whyのラベルが付けられている）
 - リンク先にWhy情報が記述されているかを調査
 - 記述されていないものはどのようなものか

RQ1 | リンク先のWhy情報の有無

- 結果

- Why情報が提供されていないもの
 - リンクURLの破損（1件）
 - コンテンツの繰り返し（59件）
 - 下手な文章（9件）
 - Whatに関する追加情報（21件）
- 先行研究でWhyを含むと誤ってラベル付けされたコミットメッセージは89件（14.57%）

RQ1の回答

コミットメッセージの中で参照されている課題報告/プルリクエストの
約15%がWhy情報を提供していない

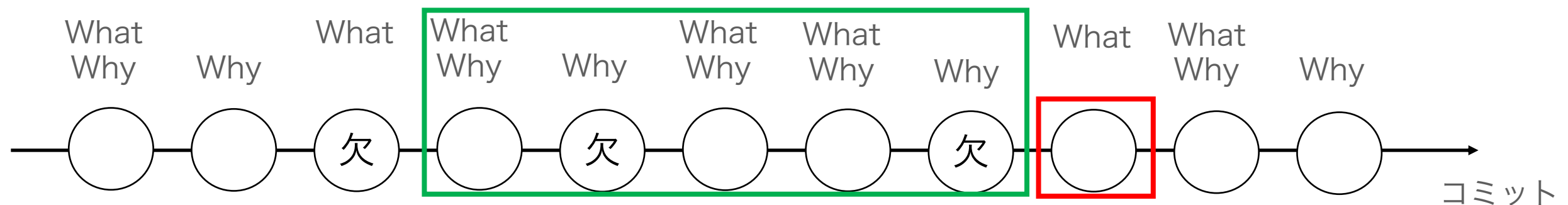
RQ2 | コミットメッセージの質の影響

- アプローチ

- Apacheの17プロジェクトから，以下の条件に合うものを除外した91,926件のコミットをデータセットとする
 - リファクタリングコミット
 - ボットがコミットメッセージを記述しているコミット
- SZZを用いてデータセットから欠陥混入コミットを特定
- 以下のウィンドウ品質スコアを計測
 - Window size = 5, 10, 20, 50, 100, 200, 400, 800, 1000

$$\text{Window Quality Score} = \frac{\# \text{Commits with positive labels in the window}}{\text{Window size}} \quad (1)$$

ウィンドウ品質スコアのイメージ



RQ2 | コミットメッセージの質の影響

- アプローチ
 - Welchのt検定^[58]とCohenのD^[23]を使用して，欠陥のあるコミットとないコミットの間のウィンドウ品質スコアの違いを分析
 - コミットレベルとファイルレベルで調査
 - コミットレベル：全てのコミットが対象
 - ファイルレベル：同じファイルを変更しているコミットが対象
 - 一般化線形回帰モデル（GLM）を構築
 - 従属変数：コミットが欠陥混入コミットかどうか
 - 独立変数：WhatとWhyのウィンドウ品質スコア，コミットメッセージの単語数
 - 3つのGLMを構築し，係数を比較

[58] G. D. Ruxton, "The unequal variance t-test is an underused alternative to student's t-test and the mann-whitney u test," Behavioral Ecology, vol. 17, no. 4, pp. 688–690, 2006.

[23] M. J. Diener, "Cohen's d," The Corsini encyclopedia of psychology, pp. 1–1, 2010.

RQ2 | コミットメッセージの質の影響

- 結果 (ウィンドウ品質スコア)

コミットレベル

Window Size	Welch's t-test p-value			Cohen's D		
	Why	What	Good	Why	What	Good
5	0.00045	9.39883e-06	0.00095	0.03168	0.03751	0.02911
10	7.21172e-07	1.22335e-09	2.59436e-06	0.04566	0.05053	0.04136
20	1.41449e-06	2.97021e-17	2.26556e-09	0.04514	0.06841	0.05272
50	3.89681e-07	4.90136e-21	5.14493e-10	0.04740	0.07620	0.05478
100	0.00155	8.71832e-18	1.41634e-06	0.02972	0.07131	0.04280
200	0.00544	1.35567e-10	2.87051e-06	0.02565	0.05459	0.04142
500	7.90503e-09	0.00282	0.00510	0.05599	0.02812	0.02603
800	1.68103e-32	1.32784e-20	1.66446e-19	0.11966	0.09197	0.08711
1000	5.11631e-47	5.25024e-33	9.56609e-33	0.14742	0.12043	0.11686

ファイルレベル

Window Size	Welch's t-test p-value			Cohen's D		
	Why	What	Good	Why	What	Good
5	2.35592e-19	3.72234e-11	1.77846e-22	0.09121	0.06700	0.10012
10	9.42900e-19	0.00079	1.13363e-14	0.09191	0.03523	0.08057
20	6.40050e-21	3.91683e-08	2.34095e-18	0.09644	0.05554	0.08948
50	1.60110e-22	4.66404e-10	2.16190e-17	0.10138	0.06378	0.08901
100	2.19140e-10	0.01863	3.63092e-08	0.07830	0.02917	0.13972
200	1.44070e-08	0.48416	0.00224	0.06268	0.00765	0.03380
500	2.32601e-19	0.02141	6.43257e-12	0.10113	0.02620	0.07755
800	1.24497e-08	0.84772	0.00049	0.06302	0.00210	0.03869
1000	1.51677e-05	0.01347	0.45646	0.04867	0.02819	0.00839

- 以前のコミットメッセージの品質は、将来のコミットの欠陥発生率に統計的に有意な影響を与える

RQ2 | コミットメッセージの質の影響

- 結果 (係数の違い)

コミットレベル

Window Size	Coefficient of Volume	Coefficient of What	Coefficient of Why
5	0.00373	0.11740	0.08787
10	0.00367	0.19327	0.17434
20	0.00363	0.34908	0.14494
50	0.00361	0.44871	0.11038
100	0.00364	0.48256	0.09678
200	0.00370	0.37022	0.13113
500	0.00383	0.22888	0.50962
800	0.00390	0.25426	0.73425
1000	0.00391	0.29600	0.83342

ファイルレベル

Window Size	Coefficient of Volume	Coefficient of What	Coefficient of Why
5	0.00393	0.24889	0.43222
10	0.00388	0.00295	0.55609
20	0.00421	0.34908	0.59306
50	0.00485	0.12108	0.64131
100	0.00366	0.16979	0.77763
200	0.00439	0.26760	0.45286
500	0.00452	0.25065	0.44335
800	0.00416	0.28544	0.45567
1000	0.00437	0.56136	0.42511

RQ2の回答

コミットメッセージに含まれるWhatとWhyの情報は、コミットメッセージの単語数と比較して、欠陥の発生しやすさとの関連性が有意に高い

RQ3 | コミットメッセージの質の進化

- アプローチ

- 以下の週品質スコアを計算（415週分）

$$\text{Weekly Quality Score} = \frac{\# \text{Commits with positive labels in a week}}{\# \text{Commits made in a week}} \quad (2)$$

- 週番号と週品質スコアのスピアマン相関分析を実施

- 結果

	係数	p値
What	- 0.79953	1.78787e-93
Why	- 0.42674	8.51816e-20
Both	- 0.80203	1.78787e-94

RQ3の回答

全体的なコミットメッセージの質は時間と共に低下する

まとめ

- 目的
 - 質の高いコミットメッセージに関する様々な調査
- RQ
 - コミットメッセージの中で参照されている課題報告/プルリクエストの約15%がWhy情報を提供していない
 - コミットメッセージに含まれるWhatとWhyの情報は、コミットメッセージの単語数と比較して、欠陥の発生しやすいとの関連性が有意に高い
 - 全体的なコミットメッセージの質は時間と共に低下する

所感

- 全体的に曖昧
 - RQ2のアプローチの解釈が合ってるのかもわからん
- RQ2の結果に違いが出るんだね. . .
 - 欠陥混入コミットはそうでないものよりも圧倒的に少ないイメージ
- 目視したデータが欲しかった. . .