

Predicting Developers' Negative Feelings about Code Review

ICSE '20, May 23–29, 2020, Seoul, Republic of Korea

紹介者 森岡勇介

背景(1)

コードレビューは欠陥を見つける、知識を共有する、保守性を高めるなどに役立つ

問題点

- 開発者同士で否定的な対人関係を持っている可能性がある
- 開発者のストレスや不満につながるかもしれない
- 先行研究では「職場での必然的なストレス」^[21]として研究されていた

[21] Scott Schieman and Sarah Reid. 2008. Job authority and interpersonal conflict in the workplace. *Work and Occupations* 35, 3 (2008), 296–326.

背景(2)と目的

レビューアが変更要求をブロックしている間のコードレビューにおける不必要な対人対立の認識を「PushBack」とする

- ほとんどはコードレビューに対して満足している
- しかし実際にこのような否定的な体験を報告している開発者がいる
- PushBackを定量的に研究した論文はまだない

目的

PushBackを定量的に測る

リサーチクエスチョン

リサーチクエスチョン

RQ1: How frequent are negative experiences with code review?

コードレビューの際にいやな経験をする頻度は？

RQ2: What factors are associated with pushback occurring?

PushBackが発生する要因は？

RQ3: What metrics detect author-perceived pushback?

(著者が認識した)PushBackを検出するメトリクスは？

アプローチ | 定性的な調査

- 定性的な調査と定量的な調査を組み合わせる
- 分析の対象はGoogle

定性的な調査

インタビュー

- PushBackの定義の改善
- 定量的な分析に使うメトリクスを抽出
- CRにPushBackを結びつける
etc...

メトリクス

Metric 1 : レビューの回数

Metric 2 : レビューにかかった時間

Metric 3 : マージされるまでにかかった時間

CRとPush Backを 結びつける

PushBackを5つのタイプに分類して
それぞれのCRにPushBackを感じるかどうか
を結びつける

PushBack

1. 対人関係の衝突
2. 受け入れられるまでの時間
3. 過度な変更要求
4. 将来のコードレビューへの不安
5. ストレス

CR : Change Request

Google内でのコードレビューの名称

アプローチ | CRの属性

CRの属性ごとにPushBackの発生率を調べて RQ2を解決するために
CRの属性を定義した

RQ2 : PushBackが発生する要因は？

CRの主な属性

- | | |
|-----------------------------------|-------------|
| 1. Readability | 読みやすさ |
| 2. CR size | コードレビューのサイズ |
| 3. New employee status, seniority | 作成者の年功序列 |
| 4. Number of Reviewers | レビューする人数 |

アプローチ | CRの分類

全体の90パーセンタイルを正常としてそれ以上を潜在的なPush Backのあるレビューとしてフラグをつけた

		CR Criteria for Inclusion		
	Surveyed CRs	Flag: Review Time	Flag: Shepherd Time	Flag: Rounds of Review
0 Flags	250			
1 Flag	200	✓		
	200		✓	
	200			✓
2 Flags	100	✓	✓	
	100	✓		✓
	100		✓	✓
3 Flags	100	✓	✓	✓

Table 1: Dist. of surveyed CRs by metric flag combinations.

潜在的なPushBackがあるとフラグをつける条件

Metric 1 : レビューの回数
9つ以上のバッチのコメント

Metric 2 : レビューにかかった時間
48分のレビュー時間

Metric 3 : マージされるまでにかかった時間
112分以上の合計時間

結果 | RQ1

RQ1 : コードレビューの際にいやな経験をする頻度は？

インタビューの時にCRを割り当てて、5つのPushBackを感じるかどうか質問した

- | | |
|---------------------|-----------------|
| 1. 対人関係の衝突を感じた | 2%($\pm 1\%$) |
| 2. 受け入れが必要以上に長いと感じた | 5%($\pm 2\%$) |
| 3. 過度な変更要求だと感じた | 2%($\pm 1\%$) |
| 4. 将来のCRに不安があると感じた | 4%($\pm 2\%$) |
| 5. レビューに不満を感じた | 3%($\pm 2\%$) |

さらに、フラグがついていないCRの 11%($\pm 3\%$)は
少なくとも一つPushBackを含むとわかった

()内 : Clopper-Pearson信頼区間

結果 | RQ2

RQ2 : PushBackが発生する要因は？

CRの属性全てを利用してロジスティック回帰を実行した

- 読みやすさ、作成者の年功序列、レビューする人数はPushBackを予測する属性ではなかった
- CRのサイズがほとんどのPushBackを予測する属性だった
非常に小さな変更に対して250～999行の変更があるCRは4.7倍将来のコードレビューの不安を抱える

結果 | RQ3

RQ3 : PushBackを検出するメトリクスは?

		CR Criteria for Inclusion		
	Surveyed CRs	Flag: Review Time	Flag: Shepherd Time	Flag: Rounds of Review
0 Flags	250			
1 Flag	200	✓		
	200		✓	
	200			✓
2 Flags	100	✓	✓	
	100	✓		✓
	100		✓	✓
3 Flags	100	✓	✓	✓

Table 1: Dist. of surveyed CRs by metric flag combinations.

Feeling of Pushback	0 Flags	3 Flags
frustration	1%	27%
interpersonal conflict	0%	22%
acceptance was withheld for too long	4%	27%
reviewer asked for excessive changes	1%	16%
feeling negatively about future CRs	4%	8%
any pushback behavior	7%	49%
2+ types of pushback behavior	1%	27%

メトリクス

Metric 1 : レビューの回数

Metric 2 : レビューにかかった時間

Metric 3 : マージされるまでにかかった時間

0フラグのものより3フラグ立っているもののCRを担当した人の方が不満を報告している

メトリクスは有効である

結果 | RQ3

RQ3 : PushBackを検出するメトリクスは?

メトリクスとCRの属性を考慮して回帰分析を行う

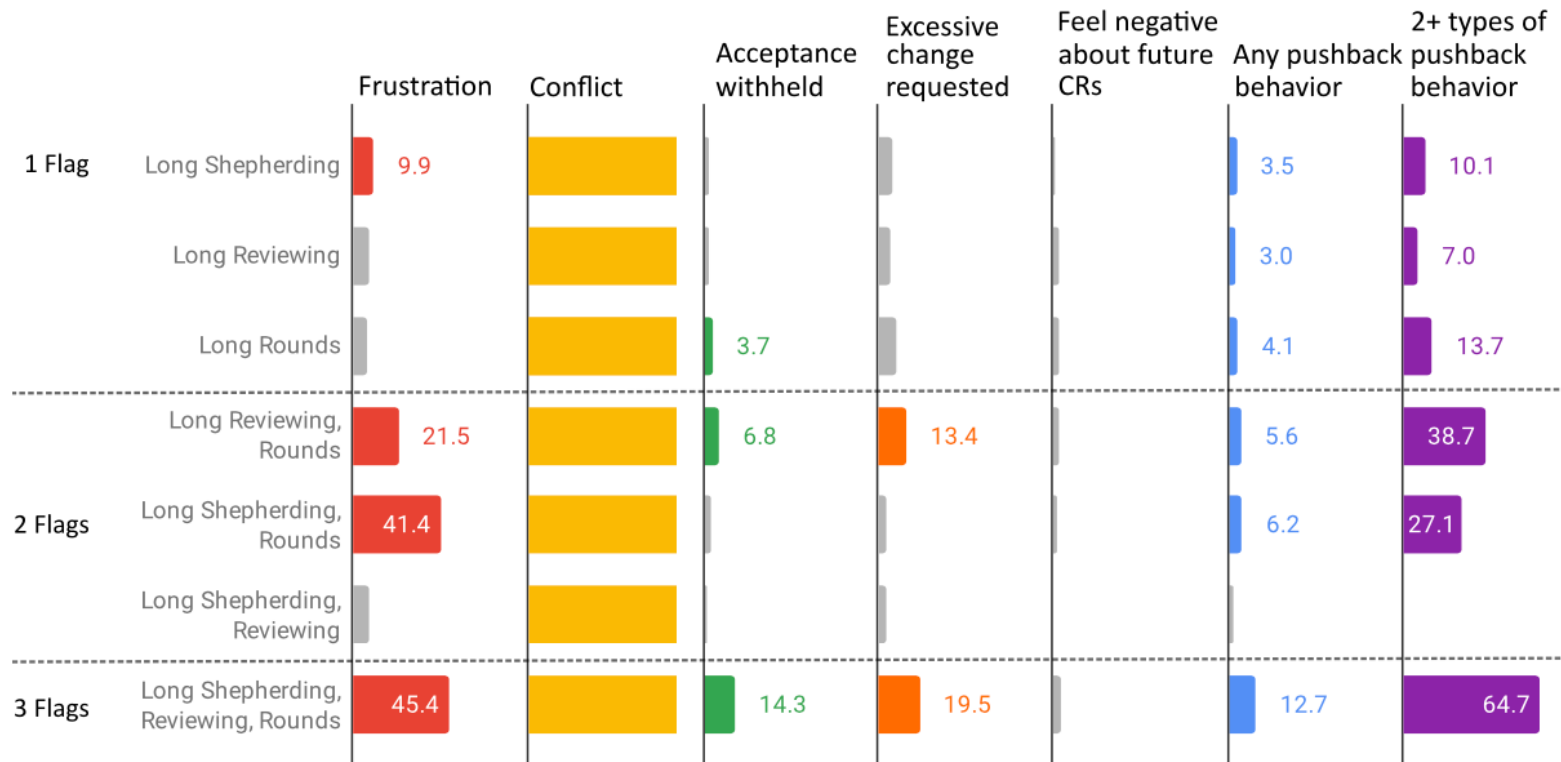
Pushback Feeling	Precision	Recall
frustration	0.11	0.98
interpersonal conflict	0.10	1.00
acceptance was withheld for too long	0.10	0.93
reviewer asked for excessive changes	0.07	0.97
feeling negatively about future CRs	0.06	0.87

Table 4: Precision and recall by feeling of pushback.

かなり高いRecallと低い
Precisionとなった

結果 | RQ3

RQ3 : PushBackを検出するメトリクスは?



レビューにかかった時間、マージされるまでの時間の二つにフラグが立っていてもレビューの回数が少ないとフラグが立っていないレビューよりPushBackをあまり感じない

結果 | RQ3

RQ3 : PushBackを検出するメトリクスは?

レビューに時間がかかってマージされるまでに時間がかかって
も回数が簡潔であればストレスを感じにくいと解釈できる

まとめと考察

- 今回考えられたメトリックでは、**recall**が高いものの**precision**が低い結果となった
- ほとんどのコードレビューには問題がないということもわかった
- しかし、問題のない(フラグが一つも立っていない)コードレビューは退屈であると報告されるケースがあった
- **Recall**が高いため、他の人に助けを求める際に現状を強調するのに適している可能性がある
- より**Pushback**が蔓延している可能性があるコミュニティを特定できる可能性がある

所感

- Recallは確かに高いが、precisionが低すぎて考察にあげた特定の使い方し
か出来ないのが残念だと感じた
- 定量的な分析が少しでほとんどが定性的な分析であったと感じた