

Code Coverage and Test Suite Effectiveness: Empirical Study with Real Bugs in Large Systems

Pavneet Singh Kochhar, Ferdian Thung, David Lo

背景

- 一般的にテストの妥当性を測るためには、カバレッジが用いられる
- しかし、カバレッジとテストスイートの質との関係は明らかになっていない

先行研究

- Gopinathら, Inozemtsevaらはコードカバレッジとテストスイートのバグ検出能力との相関を調査している
 - Randoop（テストコードを自動生成するツール）を用いてコードカバレッジを操作している

先行研究

- Gopinathら, Inozemtsevaらはコードカバレッジとテストスイートのバグ検出能力との相関を調査している
 - Randoop（テストコードを自動生成するツール）を用いてコードカバレッジを操作している

先行研究では、ミュータントバグをプロダクトコードに埋め込むことで、測定しているため実際の不具合とは異なる可能性がある

目的

- 実際のバグを用いて、テストスイートの質（どの程度バグを検知できるか）とカバレッジとの相関を測定する

リサーチクエスチョン

RQ1 テストスイートのサイズはテストの質と相関があるか？

RQ2 テストスイートのカバレッジはテストの質

前提知識

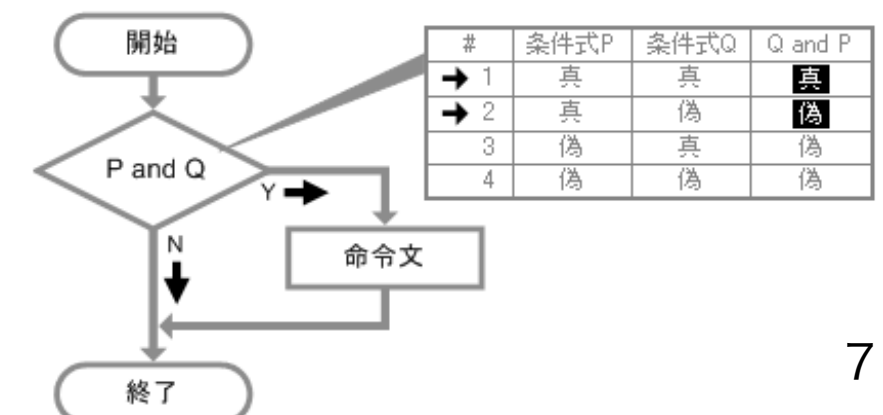
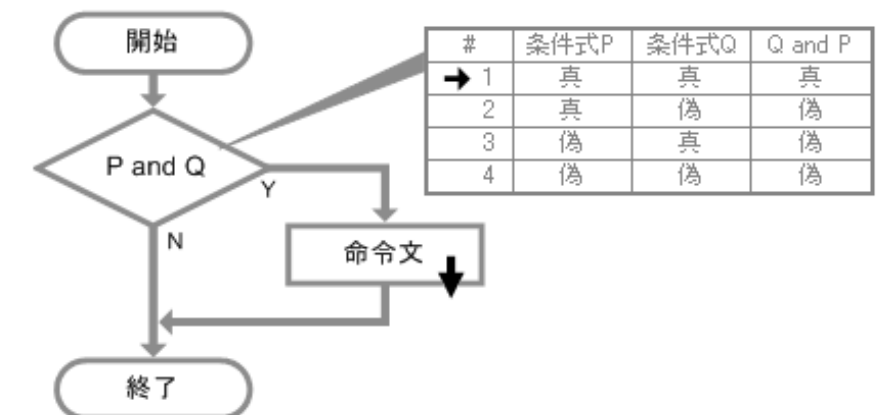
- Randoop



- Java用のテストJunit形式で自動生成するツール
- フィードバック指向ランダムテスト生成

- カバレッジ（テストの網羅率）

- ステートメントカバレッジ
 - ▶ プログラムの命令文を一回はテストする
- ブランチカバレッジ
 - ▶ プログラムの可能なすべての分岐を一回はテストする



対象プロジェクト

- HTTPClient
 - Javaライブラリ
 - クライアント側のHTTPサービスを構築する
- Rhino
 - JavascriptをJavaで実装しているオープンソースプロジェクト

TABLE I: Lines of Code of HTTPClient and Rhino

Project	Lines of Code
HTTPClient	122,288
Rhino	116,065

アプローチ概要

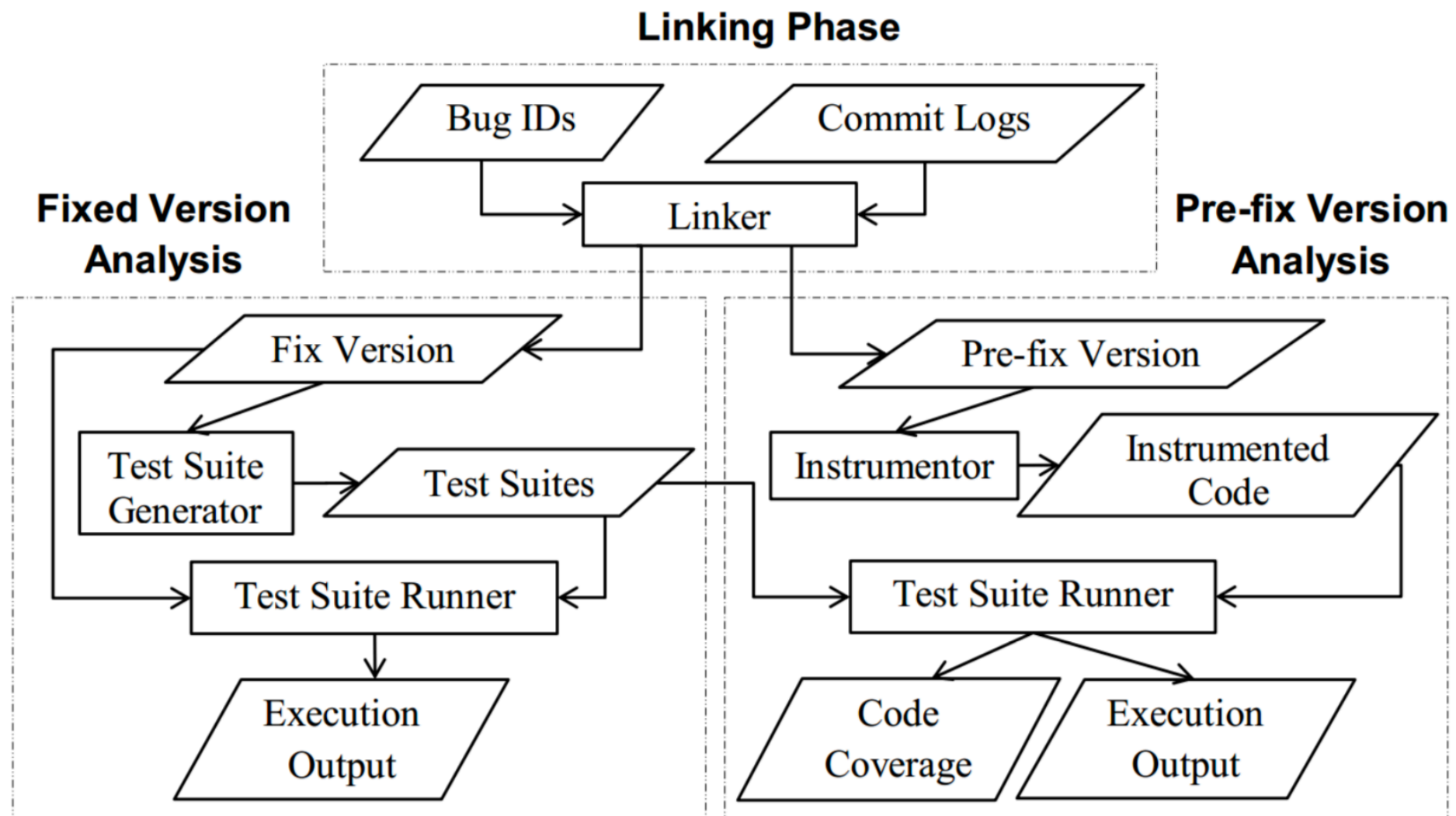


Fig. 1: Overall Process

テストサイズとテストカバレッジの調節

- Rndoopを用いて，JUnitのテストを生成する
- テストスイートのサイズとカバレッジを調整し，バグの検出数との相関を測る

TABLE II: Average Test Suite Size

Project	% of Original Test Suite Size					
	0.2	0.5	1	5	10	100
HTTPClient	7.43	15.62	39.13	197.82	396.17	3967.00
Rhino	7.64	16.01	40.10	202.52	405.46	4059.28

TABLE III: Average Coverage

Project	Coverage	% of Original Test Suite Size					
		0.2	0.5	1	5	10	100
HTTPClient	Line	7.5	11.0	17.2	28.0	31.8	37.4
	Branch	2.8	4.4	7.6	14.4	17.2	22.5
Rhino	Line	6.4	8.7	11.6	17.0	19.4	27.1
	Branch	3.0	4.2	5.8	9.0	10.5	16.5

結果：RQ 1

RQ1 テストスイートのサイズはテストの質と相関があるか？

TABLE IV: Point Biserial Correlation Values for Test Suite Size and Test Suite Effectiveness

	HTTPClient	Rhino
r_{pb}	0.70	0.37
r_{pb}^2	0.49	0.14
p-value	$4.69e^{-20}$	$4.48e^{-11}$

- 弱い相関がある

Value Range	Correlation
$r_{pb}^2 \geq 0.81$	Very strong
$0.49 \leq r_{pb}^2 < 0.81$	Strong
$0.25 \leq r_{pb}^2 < 0.49$	Moderate
$0.09 \leq r_{pb}^2 < 0.25$	Weak
$0.00 \leq r_{pb}^2 < 0.09$	Very weak

結果：RQ2

RQ2テストスイートのカバレッジはテストの質

TABLE V: Point Biserial Correlation Values for Test Suite Coverage and Test Suite Effectiveness

	Statement		Branch	
	HTTPClient	Rhino	HTTPClient	Rhino
r_{pb}	0.57	0.77	0.60	0.74
r_{pb}^2	0.33	0.59	0.36	0.55
p-value	$8.71e^{-14}$	$8.59e^{-42}$	$3.55e^{-15}$	$3.67e^{-39}$

- 強い相関がある

まとめ

- テストスイートのサイズは、テストスイートの有効性と弱い相関がある
- コードカバレッジは、テストスイートの有効性に適度または強く相関がある