How Well do Test Case Prioritization Techniques Support Statistical Fault Localization
- An empirical study on the effectiveness issue of continuous integration

Bo Jiang, Zhenyu Zhang, T.H. Tse, and T.Y. Chen

The University of Hong Kong
Hong Kong
Swinburne University of Technology
Melbourne, Australia
Contents

- Introduction and background
- Problems and research questions
- The empirical study
- Results and analysis
- Related work
- Conclusion
Contents

- Introduction and background
- Problems and research questions
- The empirical study
- Results and analysis
- Related work
- Conclusion
Motivating Example

- A previous software development scenario
  - Source code version control

![Diagram]

- Fix a bug
- Get revision suggestion
- Add in some feature
- See whether it works
- Calibrate UI
- Check the effect
Motivating Example

- A previous software development scenario
  - Source code version control

- Fix a bug
- Get revision suggestion
- Add in some feature
- See whether it works
- Calibrate UI
- Check the effect

Fixed. Compiles. But any drawback? All crucial functionality? Smoke test?
Motivating Example

- A modern software development scenario
  - More than a source code version control

Fix a bug
Get revision suggestion
Add in some feature
See whether it works
Calibrate UI
Check the effect
A Practice: Continuous Integration

- *Continuous Integration* (CI) [1] is such a practice.
A CI framework

Developers submit to CI Server

Simultaneously frequent submit
A CI framework

submit

CI Server

feedback

Feedback (build report)
A CI framework

Developer

submit

CI Server

feedback

A FAST
commit build

Test Cases

execute

Execution Information
A CI framework

A fraction of test cases, for a FAST build

Developer

CI Server

submit

feedback

select

execute

Test Cases

Execution Information
A CI framework

Developer

submit

CI Server

feedback

Test Cases

prioritize

Test Case Prioritization technique used

Execution Information

Test Cases

select

execute
A CI framework

Test Cases

prioritize

select

execute

Automatic Fault Localization technique used

Execution Information

Suspicious List

generate

Test Cases

submit

CI Server

Developer

feedback

predict
A CI framework

Test Cases

prioritize

select

execute

Test Cases

Any problem?

Suspicious List

submit

CI Server

feedback

predict

Execution Information

submit

CI Server
Contents

- Introduction and background
- **Problems and research questions**
- The empirical study
- Results and analysis
- Related work
- Conclusion
A Dilemma Existing in CI:

- Fast commit build needs
  - **Less** test cases (shorten the response time of CI cycle)
  - The less, the more efficient and fast

- Automatic Fault localization
  - **More** test cases (provide more information)
  - The more, the more effective and accurate
Research Questions:

- RQ1: To what extent will a fault localization technique be affected, if only a fraction of high-priority test cases are used as input?

- Any other question?
A CI framework

Test Cases

prioritize

select

execute

Data dependence

Execution Information

submit

feedback

Suspicious List

CI Server

Developer
More Related Issues of CI:

- Different aims
  - Test Case Prioritization (TCP) techniques
    - E.g., to increase the rate of failure detection
  - Fault Localization (FL) techniques
    - E.g., to predict suspicious program location

- Since TCP affects FL, can TCP both detect failures earlier and effectively support FL?
More Research Questions:

- RQ1: To what extent will a fault localization technique be affected if only a fraction of high-priority test cases are used as input?
- RQ2: With a view to fasten the localization of fault, is there any particularly outstanding strategy for TCP?
- RQ3: Is random TCP an acceptable cost-effect TCP strategy?
Contents

- Introduction and background
- Problems and research questions
- The empirical study
- Results and analysis
- Related work
- Conclusion
Subject Programs

- The Siemens suite programs
- The Software-artifact Infrastructure Repository (SIR)

<table>
<thead>
<tr>
<th>Program</th>
<th># of faulty versions</th>
<th>Executable LoC</th>
<th>Test Pool Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>tcas</td>
<td>41</td>
<td>133–137</td>
<td>1608</td>
</tr>
<tr>
<td>schedule</td>
<td>9</td>
<td>291–294</td>
<td>2650</td>
</tr>
<tr>
<td>schedule2</td>
<td>10</td>
<td>261–263</td>
<td>2710</td>
</tr>
<tr>
<td>tot_info</td>
<td>23</td>
<td>272–274</td>
<td>1052</td>
</tr>
<tr>
<td>print_tokens</td>
<td>7</td>
<td>341–342</td>
<td>4130</td>
</tr>
<tr>
<td>print_tokens2</td>
<td>10</td>
<td>350–354</td>
<td>4115</td>
</tr>
<tr>
<td>replace</td>
<td>32</td>
<td>508–515</td>
<td>5542</td>
</tr>
</tbody>
</table>
Experiment Setup

S0: For each faulty program
   (121 faulty versions in total)

S1: Randomly select test cases from test pool to form a test suite
   (Iterate different test suite sizes (50/100/200/300/400/500))

S2: Use a TCP technique to prioritize the test cases in the test suite
   (Iterate 9 TCP techniques)

S3: Execute program over a fraction of high-priority test cases
   (Iterate different fractions (10%/30%/50%/70%/90%/100%))

S4: Apply a FL technique, generate a suspicious list, and evaluate its effectiveness
   (Iterate 4 FL techniques)

S5: Repeat S1 to S4 for 100 times to reduce the affection of noise
Selection of Techniques

- 4 FL techniques
  - Tarantula [2]
    \[ \text{suspiciousness}_{T}(s) = \frac{\% \text{failed}(s)}{\% \text{passed}(s) + \% \text{failed}(s)} \]
  - SBI [2]
    \[ \text{suspiciousness}_{S}(p) = \frac{\text{failed}(p)}{\text{passed}(p) + \text{failed}(p)} \]
  - Jaccard [3]
    \[ \text{suspiciousness}_{O}(s) = \frac{\text{failed}(s)}{\sqrt{\text{total failed} \times (\text{failed}(s) + \text{passed}(s))}} \]
  - Ochiai [3]
Selection of Techniques

- 9 TCP techniques
  - Granularity – at what level of unit
  - Information used – based on what information
  - Strategy – what concrete strategy
    - Total [5], Additional [5], Count Metric [6], or Proportional Binary Metric [6]

<table>
<thead>
<tr>
<th></th>
<th>Statement-level</th>
<th>Function-level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coverage-based</td>
<td>“Total” strategy</td>
<td>TS</td>
</tr>
<tr>
<td></td>
<td>“Additional” strategy</td>
<td>AS</td>
</tr>
<tr>
<td>Distribution-based</td>
<td>“Count Metric” distance</td>
<td>CS</td>
</tr>
<tr>
<td></td>
<td>“Proportional Binary metric” distance</td>
<td>PBS</td>
</tr>
<tr>
<td>Random strategy</td>
<td></td>
<td>R</td>
</tr>
</tbody>
</table>
Contents

- Introduction and background
- Problems and research questions
- The empirical study
- Results and analysis
- Related work
- Conclusion
The nine points stand for effectiveness w.r.t. CF, PBF, CS, PBS, AF, AS, R, TF, and TS.
Observations – Answering RQ1

*Expense*

\[
= \frac{\text{rank of faulty statements}}{\text{number of all statements}} \quad [6]
\]

- *The lower the better*

- Example: A program consists of statements \( s_1 - s_{50} \), where \( s_{20} \) is faulty:
  - For list \(<s_{13}, s_{31}, s_{25}, s_4, s_{20}, \ldots>\)
  - Expense = \( \frac{5}{50} = 0.1 \)
Observations – Answering RQ1

Different TCP techniques have different impact on FL techniques.
Observations – Answering RQ1

Large fraction, more effective.
Observations – Answering RQ1

No significant differences for effectiveness w.r.t. 50% fraction and 100% fraction.
## Observations – Answering RQ2

<table>
<thead>
<tr>
<th>Distribution-based</th>
<th>m</th>
<th>10</th>
<th>30</th>
<th>50</th>
<th>70</th>
<th>90</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>CF</td>
<td>164%</td>
<td>86%</td>
<td>39%</td>
<td>20%</td>
<td>7%</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td>PBF</td>
<td>177%</td>
<td>102%</td>
<td>59%</td>
<td>33%</td>
<td>9%</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td>CS</td>
<td>164%</td>
<td>86%</td>
<td>41%</td>
<td>21%</td>
<td>7%</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td>PBS</td>
<td>176%</td>
<td>102%</td>
<td>59%</td>
<td>33%</td>
<td>9%</td>
<td>0%</td>
<td></td>
</tr>
</tbody>
</table>

| Random-based       | R   | 140%| 70% | 39% | 19% | 7%  | 0%  |

<table>
<thead>
<tr>
<th>Coverage-based</th>
<th>AF</th>
<th>140%</th>
<th>73%</th>
<th>43%</th>
<th>22%</th>
<th>7%</th>
<th>0%</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS</td>
<td>139%</td>
<td>64%</td>
<td>36%</td>
<td>17%</td>
<td>5%</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td>TF</td>
<td>128%</td>
<td>70%</td>
<td>45%</td>
<td>26%</td>
<td>8%</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td>TS</td>
<td>139%</td>
<td>66%</td>
<td>45%</td>
<td>22%</td>
<td>8%</td>
<td>0%</td>
<td></td>
</tr>
</tbody>
</table>

\[
\text{Relative Additional Expense} = \frac{\text{Expense}(m) - \text{Expense}(100)}{\text{Expense}(100)}
\]

- *The lower the better*
# Observations – Answering RQ2

<table>
<thead>
<tr>
<th></th>
<th>m</th>
<th>10</th>
<th>30</th>
<th>50</th>
<th>70</th>
<th>90</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Distribution-based</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PBF</td>
<td>177%</td>
<td>102%</td>
<td>59%</td>
<td>33%</td>
<td>9%</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td>CS</td>
<td>164%</td>
<td>86%</td>
<td>41%</td>
<td>21%</td>
<td>7%</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td>PBS</td>
<td>176%</td>
<td>102%</td>
<td>59%</td>
<td>33%</td>
<td>9%</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td><strong>Random</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>140%</td>
<td>70%</td>
<td>39%</td>
<td>19%</td>
<td>7%</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td><strong>Coverage-based</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AF</td>
<td>140%</td>
<td>73%</td>
<td>43%</td>
<td>22%</td>
<td>7%</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td>AS</td>
<td>139%</td>
<td>64%</td>
<td>36%</td>
<td>17%</td>
<td>5%</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td>TF</td>
<td>128%</td>
<td>70%</td>
<td>45%</td>
<td>26%</td>
<td>8%</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td>TS</td>
<td>139%</td>
<td>66%</td>
<td>45%</td>
<td>22%</td>
<td>8%</td>
<td>0%</td>
<td></td>
</tr>
</tbody>
</table>

PBF and PBS are most affected by fraction. (low robustness) (more sensitive)
### Observations – Answering RQ2

<table>
<thead>
<tr>
<th></th>
<th>m</th>
<th>10</th>
<th>30</th>
<th>50</th>
<th>70</th>
<th>90</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Distribution-based</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$CF$</td>
<td>164%</td>
<td>86%</td>
<td>39%</td>
<td>20%</td>
<td>7%</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td>$PBF$</td>
<td>177%</td>
<td>102%</td>
<td>59%</td>
<td>33%</td>
<td>9%</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td>$CS$</td>
<td>164%</td>
<td>86%</td>
<td>41%</td>
<td>21%</td>
<td>7%</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td>$PBS$</td>
<td>176%</td>
<td>102%</td>
<td>59%</td>
<td>33%</td>
<td>9%</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td><strong>Random</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R$</td>
<td>140%</td>
<td>70%</td>
<td>39%</td>
<td>19%</td>
<td>7%</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td><strong>Coverage-based</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$AF$</td>
<td>140%</td>
<td>73%</td>
<td>43%</td>
<td>22%</td>
<td>7%</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td>$AS$</td>
<td><strong>139%</strong></td>
<td><strong>64%</strong></td>
<td><strong>36%</strong></td>
<td><strong>17%</strong></td>
<td><strong>5%</strong></td>
<td><strong>0%</strong></td>
<td></td>
</tr>
<tr>
<td>$TF$</td>
<td>128%</td>
<td>70%</td>
<td>45%</td>
<td>26%</td>
<td>8%</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td>$TS$</td>
<td>139%</td>
<td>66%</td>
<td>45%</td>
<td>22%</td>
<td>8%</td>
<td>0%</td>
<td></td>
</tr>
</tbody>
</table>

AS is least affected by fraction. (high robustness) (less sensitive)
# Observations – Answering RQ3

<table>
<thead>
<tr>
<th>Distribution-based</th>
<th>$m$</th>
<th>10</th>
<th>30</th>
<th>50</th>
<th>70</th>
<th>90</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>$CF$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$PBF$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$CS$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$PBS$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random</th>
<th>$R$</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$AF$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$AS$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$TF$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$TS$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Relative Expense

\[
\frac{\text{Expense}}{\text{Expense of Random}} \quad \text{The lower the better}
\]

Expense $= 33.8\%$

Expense $= 27\%$

$33.8 / 25 = 1.250$
## Observations – Answering RQ3

<table>
<thead>
<tr>
<th></th>
<th>m</th>
<th>10</th>
<th>30</th>
<th>50</th>
<th>70</th>
<th>90</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Distribution-based</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CF</td>
<td>1.168</td>
<td>1.222</td>
<td>0.986</td>
<td>1.025</td>
<td>0.913</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>PBF</td>
<td>1.258</td>
<td>1.450</td>
<td>1.502</td>
<td>1.715</td>
<td>1.252</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>CS</td>
<td>1.172</td>
<td>1.232</td>
<td>1.029</td>
<td>1.073</td>
<td>0.947</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>PBS</td>
<td>1.255</td>
<td>1.449</td>
<td>1.501</td>
<td>1.713</td>
<td>1.250</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td><strong>Random</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td><strong>Coverage-based</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AF</td>
<td>0.997</td>
<td>1.043</td>
<td>1.084</td>
<td>1.126</td>
<td>0.887</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>AS</td>
<td>0.994</td>
<td>0.911</td>
<td>0.908</td>
<td>0.907</td>
<td>0.624</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>TF</td>
<td>0.913</td>
<td>0.996</td>
<td>1.129</td>
<td>1.326</td>
<td>1.048</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>TS</td>
<td>0.990</td>
<td>0.947</td>
<td>1.133</td>
<td>1.149</td>
<td>1.077</td>
<td>1.000</td>
<td></td>
</tr>
</tbody>
</table>

Coverage-based techniques perform better than Distribution-based techniques.
**Observations – Answering RQ3**

<table>
<thead>
<tr>
<th></th>
<th>10</th>
<th>30</th>
<th>50</th>
<th>70</th>
<th>90</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Distribution-based</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$CF$</td>
<td>1.168</td>
<td>1.222</td>
<td>0.986</td>
<td>1.025</td>
<td>0.913</td>
<td>1.000</td>
</tr>
<tr>
<td>$PBF$</td>
<td>1.258</td>
<td>1.450</td>
<td>1.502</td>
<td>1.715</td>
<td>1.252</td>
<td>1.000</td>
</tr>
<tr>
<td>$CS$</td>
<td>1.172</td>
<td>1.232</td>
<td>1.029</td>
<td>1.073</td>
<td>0.947</td>
<td>1.000</td>
</tr>
<tr>
<td>$PBS$</td>
<td>1.255</td>
<td>1.449</td>
<td>1.501</td>
<td>1.713</td>
<td>1.250</td>
<td>1.000</td>
</tr>
<tr>
<td><strong>Random</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R$</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td><strong>Coverage-based</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$AF$</td>
<td>0.997</td>
<td>1.043</td>
<td>1.084</td>
<td>1.126</td>
<td>0.887</td>
<td>1.000</td>
</tr>
<tr>
<td>$AS$</td>
<td>0.994</td>
<td>0.911</td>
<td>0.908</td>
<td>0.907</td>
<td>0.624</td>
<td>1.000</td>
</tr>
<tr>
<td>$TF$</td>
<td>0.913</td>
<td>0.996</td>
<td>1.129</td>
<td>1.326</td>
<td>1.048</td>
<td>1.000</td>
</tr>
<tr>
<td>$TS$</td>
<td>0.990</td>
<td>0.947</td>
<td>1.133</td>
<td>1.149</td>
<td>1.077</td>
<td>1.000</td>
</tr>
</tbody>
</table>

*R is a cost-effective choice.*
Contents

- Introduction and background
- Problems and research questions
- The empirical study
- Results and analysis
- Related work
- Conclusion
Related Work


- A general introduction of CI


- FL techniques


- FL techniques
Related Work


- TCP techniques


- TCP techniques


- Effectiveness Metric
Contents

- Introduction and background
- Problems and research questions
- The empirical study
- Results and analysis
- Related work
- Conclusion
Conclusion

- The effectiveness of a statistical FL technique is, on average, not much different between applying the entire test suite and applying the first half of the test suite as the FL’s inputs.

- Statistical FL + Coverage-based TCP is less sensitive (more robust) than Statistical FL + Distribution-based TCP.
  - Least sensitive TCP: Additional-Statement (AS)
  - No other TCP is less sensitive than the Random (R) in all sizes.

- The $R$ prioritization can be a cost-effective choice, in supporting fault localization in a test budget limited environment.
Future Work

- For what kind of program spectra in executions, will AS be a best TCP strategy, in supporting FL? Proof?
  - The process of a AS prioritization
    - The highest-priority test case is the one with maximum coverage
    - The 2\textsuperscript{nd} test case is the one taking in maximum additional coverage
    - When all statements have been covered, clear all flag and repeat
  - One Clue: AS maximizes the minimum number of times a statement is covered
  - ...

- Such a future work scientifically supports using AS in CI.
Your Comments are Welcome