A DATASET FOR
DYNAMIC DISCOVERY OF SEMANTIC CHANGES
IN VERSION CONTROLLED SOFTWARE HISTORIES

MSR
MAY 21, 2017

Chenguang Zhu    Yi Li
Julia Rubin      Marsha Chechik
Goal: Given a version-controlled software history, identify code changes related to a particular high-level functionality.

Uses:
- Assist in transferring functionalities across branches
- Produce focused pull-requests
- Locate features [SPLC’17]
- Etc.

Semantic History Slice: A (minimal) sub-sequence of a change history that preserves the functionality of interest as defined by a set of test cases.

[ASE’15, ASE’16, TSE’17]

https://github.com/Chenguang-Zhu/DoSC
CHALLENGES IN EVALUATING SEMANTIC HISTORY SLICING

- Need a significant number of well-documented functionalities
- Functionalities should be accompanied by test cases
- Need ground truth

Contribution of this work:
- 98 items of semantic change data, collected from 10 open-source Java projects.

https://github.com/Chenguang-Zhu/DoSC
DATASET CREATION

- Chose well-documenting projects
  - 10 Apache Java projects, using JIRA for issue tracking

- Chose projects with test cases
  - Functionalities committed together with a test suite

- Obtained ground truth
  - Ran a delta debugging-style partitioning algorithm, to produce 1-minimal slices
    - 1-minimal – removing a single commit makes the result invalid

https://github.com/Chenguang-Zhu/DoSC
DATA SCHEMA

Meta-data: YAML

Example: CALCITE-1168

```yaml
1  id: CALCITE-1168
2  description: Add DESCRIBE SCHEMA/DATABASE/TABLE/query
3  project:
4      name: Calcite
5      project url: https://github.com/MSR-2017(calcite)
6      issue url: https://issues.apache.org/jira/browse/CALCITE-1168
7      history start: 8ee6bfc6d
8      history end: aeb6bf14
9  test suite:
10     - "SqlParserTest.testDescribeSchema"
11     - "SqlParserTest.testDescribeTable"
12     - "SqlParserTest.testDescribeStatement"
13  history slice:
14     - "a065200a"
15     - "da875a67"
16  developer labeled commits:
17     - "a065200a"
18     - "da875a67"
```

https://github.com/Chenguang-Zhu/DoSC
HOW TO USE THE DATASET

1. Pick a functionality (e.g., **CALCITE-1168**), view the meta-data.

2. Clone the project repository to the local machine.

3. Extract all test cases defining the functionality.

   \[ \text{F1} \]
   \[ \text{TestA.testf()} \]
   \[ \text{TestB.testf()} \]
   \[ \text{TestC.testg()} \]
   ...
4. Extract the starting point and the ending point of the history segment (SHA-1-expressed commit number)

5. Using the selected test cases (Step 3) and history segment (Step 4) as input, run the history slicing/feature location tool to be evaluated

https://github.com/Chenguang-Zhu/DoSC
6. Compare the obtained result with ground truth we provide

1) Semantic History Slicing:

User’s result

2) Feature location:

User’s result

https://github.com/Chenguang-Zhu/DoSC
# Dataset Overview

<table>
<thead>
<tr>
<th>Project</th>
<th>#F</th>
<th>#R</th>
<th>Avg. Commits</th>
<th>Avg. Files</th>
<th>Avg. LOC</th>
<th>Avg. Tests</th>
<th>Avg. Slice</th>
<th>Avg. Reduce (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>commons-lang</td>
<td>20</td>
<td>4</td>
<td>334.25</td>
<td>191.75</td>
<td>17423.95</td>
<td>5.55</td>
<td>43.1</td>
<td>87.11</td>
</tr>
<tr>
<td>calcite</td>
<td>18</td>
<td>7</td>
<td>89.83</td>
<td>332.67</td>
<td>31150.78</td>
<td>3.39</td>
<td>6.61</td>
<td>92.64</td>
</tr>
<tr>
<td>maven</td>
<td>11</td>
<td>6</td>
<td>82.09</td>
<td>183.09</td>
<td>7153.27</td>
<td>2.27</td>
<td>8.18</td>
<td>89.24</td>
</tr>
<tr>
<td>commons-compress</td>
<td>9</td>
<td>2</td>
<td>155</td>
<td>156.33</td>
<td>7172.67</td>
<td>5</td>
<td>17.33</td>
<td>88.82</td>
</tr>
<tr>
<td>flume</td>
<td>9</td>
<td>3</td>
<td>104.11</td>
<td>299.33</td>
<td>21355.56</td>
<td>4</td>
<td>20.22</td>
<td>79.82</td>
</tr>
<tr>
<td>pdfbox</td>
<td>5</td>
<td>3</td>
<td>203</td>
<td>188.4</td>
<td>10184</td>
<td>6.2</td>
<td>2</td>
<td>98.7</td>
</tr>
<tr>
<td>commons-configuration</td>
<td>3</td>
<td>2</td>
<td>117.33</td>
<td>254</td>
<td>54576</td>
<td>6</td>
<td>20.67</td>
<td>65.61</td>
</tr>
<tr>
<td>commons-net</td>
<td>3</td>
<td>2</td>
<td>205</td>
<td>188.33</td>
<td>7202.33</td>
<td>6.67</td>
<td>29</td>
<td>87.05</td>
</tr>
<tr>
<td>commons-csv</td>
<td>4</td>
<td>1</td>
<td>79</td>
<td>28</td>
<td>2353</td>
<td>3.75</td>
<td>42.5</td>
<td>46.2</td>
</tr>
<tr>
<td>commons-io</td>
<td>16</td>
<td>2</td>
<td>138.25</td>
<td>158.38</td>
<td>8047.44</td>
<td>9.5</td>
<td>24.06</td>
<td>82.59</td>
</tr>
<tr>
<td><strong>Overall</strong></td>
<td>98</td>
<td>32</td>
<td>163.74</td>
<td>212.79</td>
<td>16521.01</td>
<td>5.24</td>
<td>21.66</td>
<td>86.77</td>
</tr>
</tbody>
</table>
STATUS

URL: https://github.com/Chenguang-Zhu/DoSC

Contents
• Meta-data template
• 98 pieces of data
• Our tool for obtaining 1-minimal history slice

Easy to extend
• ... e.g., by including more repositories
• ... or by adding histories containing bugs and failed test cases – for fault localization

WE LOOK FORWARD TO YOUR USES AND EXTENSIONS!!!!!
Thank you!

https://github.com/Chenguang-Zhu/DoSC
A RECENT USE CASE

FHistorian – A feature relationship analysis tool

Our Dataset

Feature Location

Intersection

Feature Graph

[SPLC’17]

https://github.com/Chenguang-Zhu/DoSC
Thank you!

Try it here!  
https://github.com/Chenguang-Zhu/DoSC
DATASET CREATION

- Project selection: 10 Apache Java projects, using JIRA for issue tracking.

- Functionality selection: Functionalities accompanied by a test suite, committed together.

- History range selection: Between two release versions. \((R_1 - R_2)\)

- Obtaining ground truth: Run a delta debugging-style partitioning algorithm, to produce the 1-minimal slice.

https://github.com/Chenguang-Zhu/DoSC