Subject: [Accurate and efficient refactoring detection in commit history - A second replication

Date: Q2 2018 – Q3 2018

We re-evaluate *RefactoringMiner* (*RMiner*) and *RefDiff*'s accuracy and running time for a sample[1] of 37 commits randomly selected from your original dataset (oracle). Table 1 describes the instrumentation applied to the study.

Type	**Description**
Code/Documentation	*RMiner* [2]and *RefDiff* [3]
System configuration	*Intel core* i7-8550U CPU@2.70 GHz, 16 GB DDR3, 2TB SSD, Linux Ubuntu 16.04
Development	Java 1.8.0_162, IDE *Eclipse* Oxygen.3a *Release* (4.7.3a), R 3.4.4 e *RStudio* 1.1.423
Measurement	Manual inspection for counting TPs (True Positive), FPs (False Positive) and FNs (False Negative), and the System.nanoTime() Java method for collecting the running time.

Table 1: Instrumentation

Table 2 shows the types of refactoring detected by *RMiner* and *RefDiff* for the sample, which contained 425 refactorings. We considered only projects with few refactorings because we needed to perform a manual validation regarding true positives (TP), false positives (FP) and false negatives (FN) for each project.

Refactoring type
Extract Superclass, *Move Class*, *Rename Class*
Extract Method, *Inline Method*, *Pull Up Method* *Rename Method*
Pull Up Field, *Push Down Field*, *Move Field*

Table 2: Types of refactoring detected by *RMiner* and *RefDiff* for the sample

Figures 1-2 describe the precision and recall obtained per refactoring type. For the sample, on average, the *RMiner*'s precision was 100.0%, while *RefDiff*'s was 98.27%. With respect to recall – on average, *RMiner* obtained 92.6% and *RefDiff*, 88.42%.

Figure 1 - Precision x Refactoring Type x Detection Tool

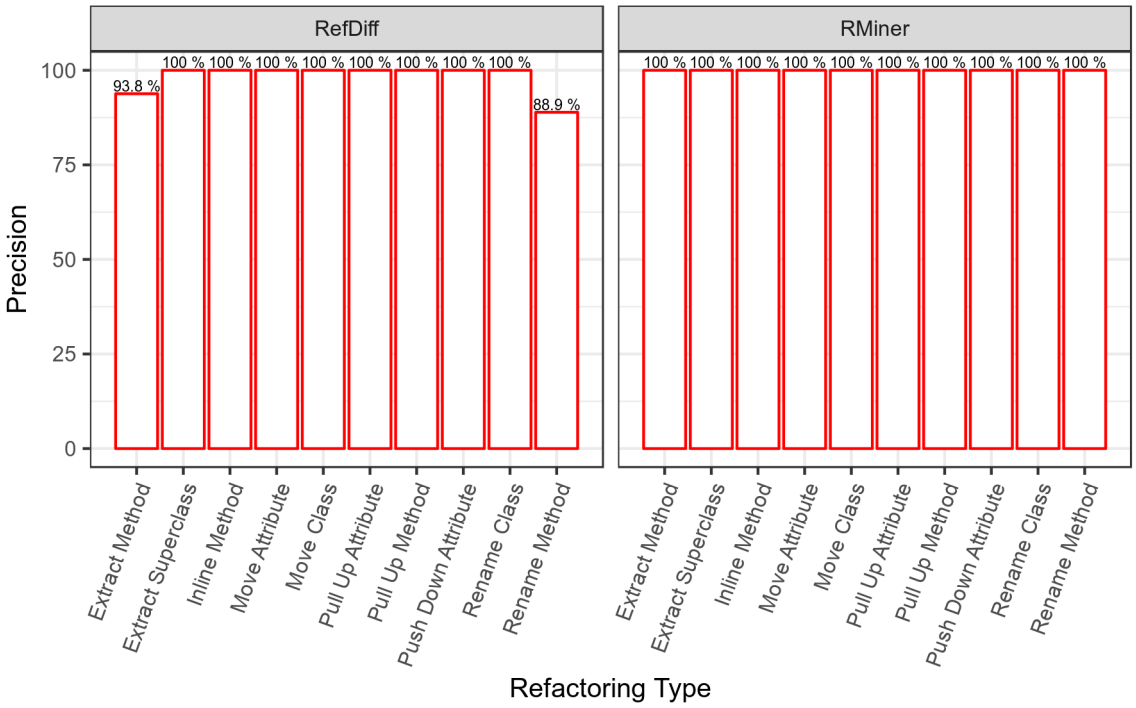
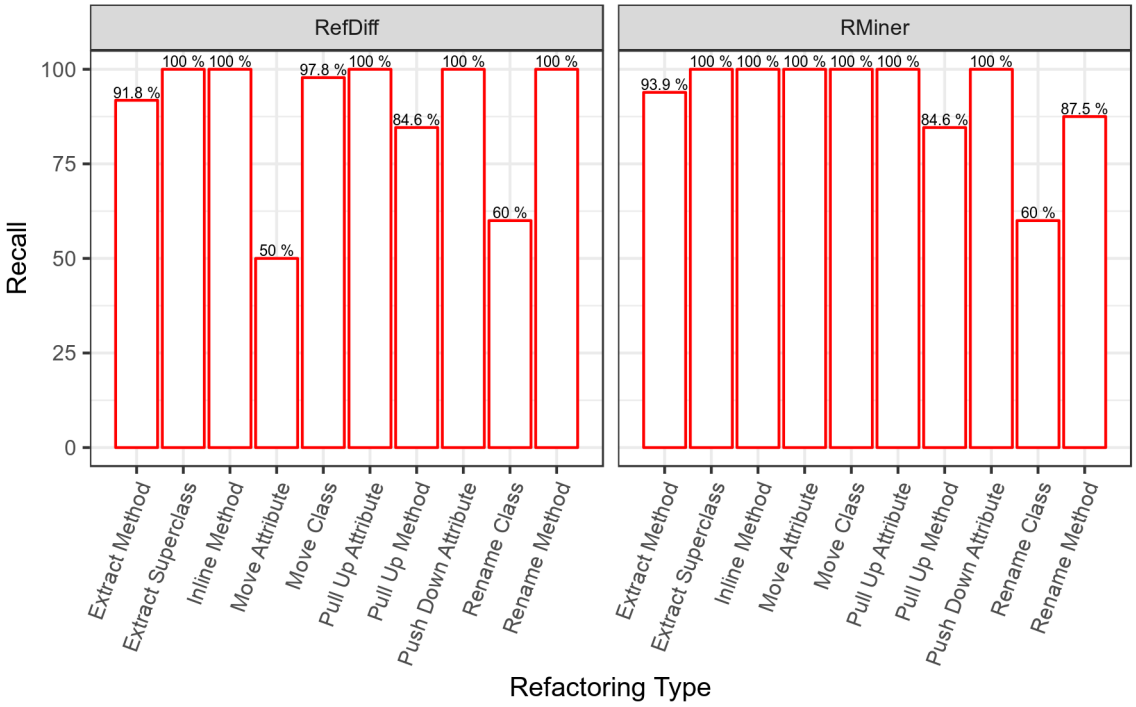
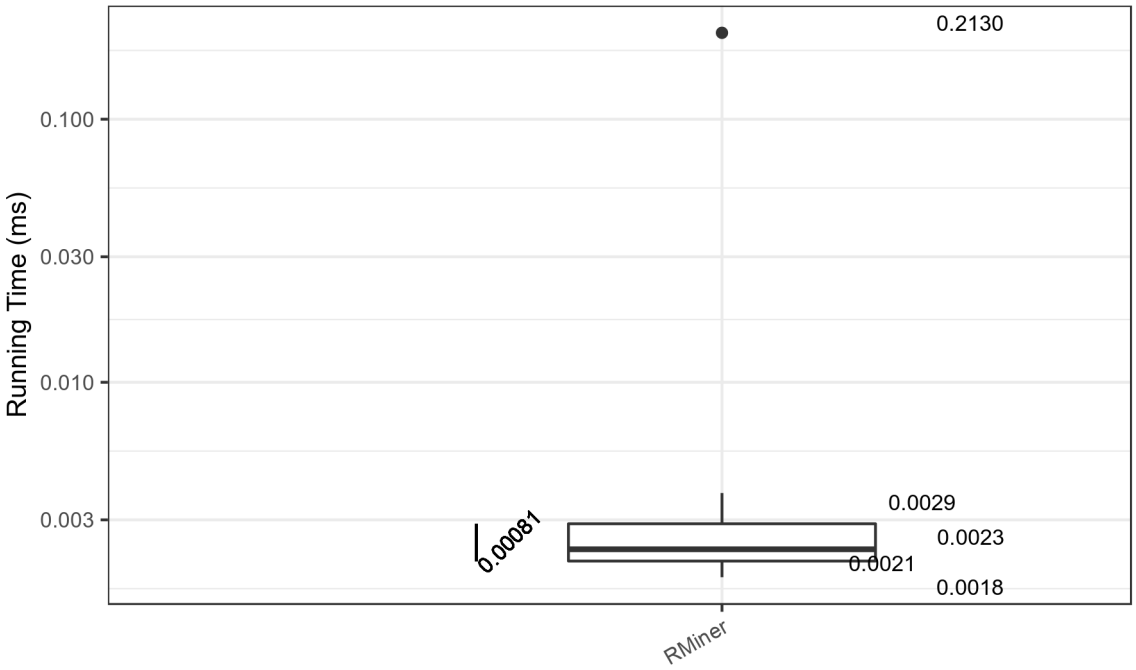


Figure 2 - Recall x Refactoring Type x Detection Tool



RMiner's running time ranged from 0.0018 to 0.2130 ms (median value of 0.0023 ms), as shown in Figure 3. Figure 4 presents the distribution for *RefDiff* running time, which ranged from 227 to 9008 ms (median value of 1522 ms).

Figure 3 - RMiner Running Time



Discussion

Our sample of 37 commits may not be considered representative due to the types of detected refactorings (Table 2).

RMiner's precision (100.0%) and recall (92.6%) were higher than *RefDiff*'s precision (98.27%) and recall (88.42%), respectively. In terms of precision and recall, we found values that go along with the ones obtained in [1].

We also applied the Wilcoxon signed rank test on the paired samples of the running time for each commit, which too rejected the null hypothesis "*RefDiff* execution time is smaller than that of *RMiner*" with a p-value <2.2e-16.

References

[1] Tsantalis, N., Mansouri, M., Eshkevari, L. M., Mazinanian, D., and Dig, D. *Accurate and efficient refactoring detection in commit history*. In Proceedings of the 40th International Conference on Software Engineering, Gothenburg, Sweden, 2018.

[1] All results for this study are available at https://github.com/flaviacoelho/rerun-replication-RMiner
[3] github.com/tsantalis/RefactoringMiner
[3] github.com/aserg-ufmg/RefDiff

Figure 4 - RefDiff Running Time