Java Bytecode Normalization for Code Similarity Analysis (Artifact)

Stefan Schott
Paderborn University, Germany

Serena Elisa Ponta
SAP Security Research, Mougins, France

Wolfram Fischer
SAP Security Research, Mougins, France

Jonas Klauke
Paderborn University, Germany

Eric Bodden
Paderborn University, Germany
Fraunhofer IEM, Paderborn, Germany

Abstract

This artifact supports the claim that different Java compilation environments can produce significantly different bytecode and that bytecode normalization applied via our tool JNORM heavily decreases the amount of compilation differences and helps boost the performance of subsequent code similarity analysis. Our artifact provides the source code of the tool JNORM and all scripts needed to reproduce the results we presented in our evaluation. Furthermore, it contains a study on the usage of different Java compilers and target levels within popular open-source projects, which showcases that the JDK compiler is by far the most relevant compiler in practice.

2012 ACM Subject Classification Software and its engineering → Compilers

Keywords and phrases Bytecode, Java Compiler, Code Similarity Analysis

Digital Object Identifier 10.4230/DARTS.10.2.20

Funding This work was partially supported by the German Research Foundation (DFG) within the Collaborative Research Centre ”On-The-Fly Computing“ (GZ: SFB 901/3) under the project number 160364472.

Related Article Stefan Schott, Serena Elisa Ponta, Wolfram Fischer, Jonas Klauke, and Eric Bodden, “Java Bytecode Normalization for Code Similarity Analysis”, in 38th European Conference on Object-Oriented Programming (ECOOP 2024), LIPIcs, Vol. 313, pp. 37:1–37:29, 2024.
`https://doi.org/10.4230/LIPIcs.ECOOP.2024.37`

Related Conference 38th European Conference on Object-Oriented Programming (ECOOP 2024), September 16–20, 2024, Vienna, Austria

Evaluation Policy The artifact has been evaluated as described in the ECOOP 2024 Call for Artifacts and the ACM Artifact Review and Badging Policy.

1 Scope

The artifact contains the source code of JNORM and all scripts that we used to obtain the claims made in the paper. This includes the scripts to reproduce the usage study on Java compilers and target levels within open-source projects (see electronic appendix), the investigation of changes induced by different JDK compiler vendors (RQ1), different JDK versions (RQ2) and different target levels (RQ3). Alongside it also contains the scripts to determine prevalence of

DAGSTUHL ARTIFACTS SERIES
Dagstuhl Artifacts Series
Schloss Dagstuhl – Leibniz-Zentrum für Informatik,
Dagstuhl Publishing, Germany

each normalization transformation (RQ5). Furthermore, it contains the extended NiCad version to support the claim that bytecode normalization significantly increases the performance of code similarity analysis (RQ5).

2 Content

The artifact package includes (within the VM image):

- jNorm tool and source code
- Fine-grained paper results
- Evaluation pipeline
- Jimple-Extended NiCad version
- Compiler-usage and target/source level evaluation scripts
- Electronic appendix containing a usage study of Java compiler and target levels in open-source Java projects (outside VM image)

3 Getting the artifact

The artifact endorsed by the Artifact Evaluation Committee is available free of charge on the Dagstuhl Research Online Publication Server (DROPS). In addition, the artifact is also available at: `https://doi.org/10.5281/zenodo.12625104`.
Furthermore, jNorm is available at: `https://github.com/stschott/jnorm-tool`.

4 Tested platforms

The artifact is hosted inside an Ubuntu VM. We tested the artifact on VirtualBox 7.0 running on a Windows 11 machine with an AMD Ryzen 5 3600 (6 Core, 2.6 GHz) CPU and 32 GB main memory.

5 License

All content specifically created for the artifact is available under CC-BY license.

6 MD5 sum of the artifact

6eab9bda294f45b71820d0c87286e536

7 Size of the artifact

9.84 GiB

A Getting started

At first the VM image needs to be imported into a VM software of the reviewers choosing. We used VirtualBox. To login into the VM the following credentials can be used:

- **User:** jnorm
- **Password:** jnorm

Once logged in, the artifact is located on the Desktop within the `artifact` folder. (Specificially under the path `/home/jnorm/Desktop/artifact`).

There are 4 scripts located in this directory. To run a scaled down version of the primary evaluation run the following command within the `/home/jnorm/Desktop/artifact` directory: `./sample_evaluation.sh`.
To run the full evaluation run the following command: `./full_evaluation.sh`.