Vrije Universiteit Brussel

Proposal bachelor thesis

Title: Mining the Maven Repository for Cross-System Patches

Promotor: Coen De Roover

Includes preparation course: Yes

Context

Maven is one of the most popular repositories for open-source Java libraries. Several versions of their source and compiled code are hosted there. The SOFT laboratory has access to a huge mirror of this repository in Japan that is several gigabytes in size.This enables you to mine the repository for insights into whether and how bug fixes and feature enhancements for one library are ported to or percolate to different libraries in practice. In turn, these insights will enable us to build more effective tools that recommend developers how to perform such ports themselves.

Proposal bachelor thesis

The goal of the thesis is to gather these insights using a research method that is similar to the one proposed in the paper[1] "*Identifying Source Code Reuse across Repositories using LCS-based Source Code Similarity"* by Kawamitsu et al. and to the one proposed in the paper[2] "*A Case Study of Cross-System Porting in Forked Projects*" by Ray et al.:

1) Recreate the genealogy of the involved libraries based on the similarity and age of their files (i.e., from which version of a library was code copied to which version of another library).
2) Compute the differences between successive versions of each system.
3) Find similar differences across all system versions as an indication of which bug fixes and feature enhancements make it from one system to another.

While off-the-shelf tools and algorithms can be used to gather the data, the challenge lies in the massive size of the Maven repository, which will require some clever (possibly parallel) orchestration and storage.

---

[1] http://soft.vub.ac.be/Publications/2014/vub-soft-tr-14-10.pdf
[2] http://users.ece.utexas.edu/~miryung/Publications/fse2012-porting.pdf

Preparatory course bachelor thesis

To prepare for this thesis, you will perform a literature study on the state of the art in mining software repositories. The focus will be on similarity metrics for source code, and on techniques for reconstructing software genealogies.