Thesis for the Degree of Doctor of Philosophy

Empowering Empirical Research in Software Design: Construction and Studies on a Large-Scale Corpus of UML Models

Truong Ho-Quang

Presentation:
October 9th, 2019, 13:00
Dome of Visions
Lindholmsplatsen, 41756 Gothenburg

Opponent:
Dr. Klaas-Jan Stol (University College Cork, Ireland)

Grading Committee Members:
Dr. Maria Teresa Baldassarre (University of Bari Aldo Mori, Italy)
Dr. Christoph Treude (University of Adelaide, Australia)
Dr. Sebastian Herold (Karlstad University, Sweden)

Supervisors:
Prof. Michel R.V. Chaudron (Main supervisor)
Regina Hebig (Co-supervisor)

The thesis is available at:
Department of Computer Science & Engineering
Chalmers University of Technology and University of Gothenburg
Gothenburg, Sweden, 2019

Phone: 031 772 6174

Abstract

Context: In modern software development, software modeling is considered to be an essential part of the software architecture and design activities. The Unified Modeling Language (UML) has become the de facto standard for software modeling in industry. Surprisingly, there are only a few empirical studies on the practices and impacts of UML modeling in software development. This is mainly due to the lack of empirical data on real-life software systems that use UML modeling.

Objective: This PhD thesis contributes to this matter by describing a method to build and curate a big corpus of open-source-software (OSS) projects that contain UML models. Subsequently, this thesis offers observations on the practices and impacts of using UML modeling in these OSS projects.

Method: We combine techniques from repository mining and image classification in order to successfully identify more than 24.000 open source projects on GitHub that together contain more than 93.000 UML models. Machine learning techniques are also used to enrich the corpus with annotations. Finally, various empirical studies, including a case study, a user study, a large-scale survey and an experiment, have been carried out across this set of projects.

Result: The results show that UML is generally perceived to be helpful to new contributors. The most important motivation for using UML seems to be to facilitate collaboration. In particular, teams use UML during communication and planning of joint implementation efforts. Our study also shows that the use of UML modeling has a positive impact on software quality, i.e. it correlates with lower defect proneness. Further, we find out that visualisation of design concepts, such as class role-stereotypes, helps developers to perform better in software comprehension tasks.

Keywords

Software Modeling, Software Design, Empirical Research, UML, Modeling Practice, Impacts of Modeling, Open Source System, Mining Software Repository, Data Mining, Data Curation, GitHub.