VUB SCIENCES & BIOENGINEERING SCIENCES

The Research Group
Software Languages Lab

has the honor to invite you to the public defence of the PhD thesis of

Camilo Velázquez-Rodríguez

to obtain the degree of Doctor of Sciences

Title of the PhD thesis:

Extracting Library Features from Incomplete Code on Stack Overflow

Promotor:
Prof. dr. Coen De Roover

The defence will take place on

Friday, March 29, 2024 at 5:00 p.m. in auditorium I.2.02

The defence can also be followed through a live stream: https://msteams.link/J84H

Members of the jury

Prof. dr. Viviane Jonckers (VUB, chair)
Prof. dr. Wolfgang De Meuter (VUB, secretary)
Prof. dr. Ann Nowé (VUB)
Prof. dr. Sam Verboven (VUB)
Prof. dr. Davide Di Ruscio, Università degli Studi dell'Aquila, Italy)
Prof. dr. Bin Lin, Radboud Universiteit, The Netherlands)

Curriculum vitae

Camilo Velázquez-Rodríguez obtained his M.Sc. in Applied Mathematics and Informatics for Administration at the Universidad de Holguín in 2016. He started his PhD at the Software Languages Lab (SOFT) in 2018 supported by the Excellence of Science Research Project SECO-ASSIST.
His research focused on designing techniques and tools to assist developers in selecting a library in a vast software ecosystem through the extraction of features. Camilo's research resulted in five publications in international peer-reviewed scientific journals and conferences, out of which one was honoured with a distinguished paper award. Camilo presented his research at international conferences and workshops. He has also co-supervised various Bachelor's and Master's theses.

Abstract of the PhD research

It is common in contemporary software development to reuse features provided by third-party libraries. Reusing features instead of re-implementing them from scratch can reduce development time and may improve the overall system quality. However, selecting an appropriate library to reuse features from can be difficult for developers due to the lack of automated tool support. Some library indices propose rankings of libraries, but these are biased towards the number of library downloads, attributed stars, etc. This may lead developers to select the most popular library instead of the library with the feature they were hoping to reuse.

This thesis makes three contributions. The first, called LiFUSO, is an automated approach to enumerating and describing the features provided by a library based on publicly available information on social coding platforms such as Stack Overflow (SO) and GitHub. LiFUSO analyses library usages within SO posts and extracts usage patterns indicative of library features. To this end, it considers both the code snippets and the surrounding natural language within each SO post that discusses the library. RESICO, the second contribution of the thesis, is an automated approach to resolving API type references within a code snippet to their corresponding fully-qualified name. As a learning-based text classification approach, RESICO needs to be trained on a corpus of programs for which a compiler has determined the correct type information. Once trained, it can take syntactically incorrect code snippets as input. The final contribution combines LiFUSO and RESICO so the former is no longer limited in scope to SO posts that have been tagged with a library's name but can also extract information from posts in which it has recognised many library types. We evaluate the impact of broadening the scope of the analysis on the quantity and quality of the uncovered library features.

The contributions are relevant to the software engineering community at large. RESICO's type resolution can be adopted by tools that need to analyse potentially incomplete code snippets, or by tools that need to determine the libraries used within a code snippet ---just like in our third contribution. LiFUSO paves the way for tool support for selecting a library from many alternatives. Finally, these contributions help understand the benefits and drawbacks of data-centric over algorithmic-centric solutions to software engineering problems.