

Research Report 2020

Administrative part

- Researcher name: PHAM Hoang Son
- Date of first research contract in INGI: 5/2/2018
- List of past contracts in INGI (if any): none
- Research advisor name(s): Prof. Kim Mens and Prof. Siegfried Nijssen
- Current funding source: Brussels Region (INTiMALS project)
- Research topic: to research and deploy novel pattern mining algorithms to discover structural regularities in source code repositories.
- PhD admission: not applicable (already postdoc)
- PhD confirmation: not applicable (already postdoc)

Teaching tasks in 2019-2020

- List of courses (1st, 2nd semester): 1 course (Databases-LINGI2172) in the 2nd semester.
- Estimated time spent on the teaching tasks (in percentage of full-time, averaged over 12 months): approximately 20% of my working time (in the period of teaching time). This amount of time includes practice section (2 hours per week) and exercise correction.

Research activities in 2019-2020

- Research topic: we focus on researching and deploying novel pattern mining algorithms to source code repositories. In particular, the first stage of the project is aim to mine for regularities in software source code. In the second stage of the project, we applied graph mining algorithms to discover library usage in software source code. These usage patterns are invaluable in understanding how the API is typically used by developers and help highlight anomalous usage. In the third stage of the project, we focus on mining changed patterns from two versions of a system. These patterns help developers to understand what pieces of source code changed from old version to new version of a system and vice versa.
- Results:
 - For the first stage, we developed a new algorithm, FREQTALS, to discover regularities in software source code. This algorithm has been applied to mine for patterns in Java and Cobol systems.
 - For the second stage of the project, we proposed to use SUBDUE, a graph mining algorithm to discover API usages from software source code. Our preliminary experimental results show that SUBDUE was not very scalable to mine usage patterns in source code. Because the size of graphs produced from software source code is often large the cost of computing subgraph isomorphisms is high.
 - For the third stage of the project, we adapted FREQTALS, a frequent tree mining algorithm which was developed in the first stage of the project, to mine for changed patterns from two versions of a system. Preliminary experimental results show that the adapted algorithm efficiently mines for changed patterns from two versions of Java software projects. It was also able to mine changed patterns in Python source codes.

- Possible difficulties:
 - The most challenging task of our current work is how to efficiently mine for patterns from large datasets. Since the input data is often very large, the mining algorithms could take a very long time to finish the search.
- Perspectives for 2020-2021:
 - Continue improving our FREQTALS algorithm for software source code mining.
 - Apply FREQTALS to mine for patterns in source code of students' exams. The purpose is to find good and bad code structures which could be useful to build a recommendation system.
 - Build a graphical tool to help users inspect discovered patterns.
 - Write papers for the current research.

Publications

- Publications and research reports (since January 1, 2019): we made effort to submit papers to conferences and programming-journal, but they were not successful.
- Future publications: our plan is to conduct more empirical analysis on a large number of software source code repositories to write a journal paper that discusses the specific pattern mining approach for source code repositories. In addition, we plan to write a conference paper that discusses specific techniques used to mine for changed patterns in software source code.