

Software Development Difficulties

Calvin Loncaric

January 11, 2016

Problem 1: Avoiding ¹unnecessary ²recomputation. Recently I found myself implementing a tool to process each one of about five hundred files. Even when run in parallel on `recycle.cs`, the entire process still took fifteen minutes. The tool was (intentionally) incomplete, and each time I run it I would get a histogram of missing features that caused failures on some files. I would implement those features, and then run the tool again to discover new problems, and so on in a loop until the tool ran on every file.

Most of the changes I made to the tool did not affect the output on most of the files, but the makefile I had in place was not smart enough to realize this fact. It would have been very difficult for me to write additional code to identify (1) which functionality I had changed in the tool and (2) which features in each file this would affect. However, I think this would have been possible for a computer to do; both the tool and each file were structured around the possible attributes a file might have. Avoiding the recomputation each time the tool ran could have saved many minutes of waiting for the entire process to complete.

Problem 2: Reading method implementations. As a graduate student this has been less important, but when I was working in industry I spent much ³more time reading code than I did writing it. In older code, it was frequently the case that the important classes and methods had many cross-cutting concerns: they would do some work, do some logging, do some cleanup, and so on. The additional work done by logging and so on obscures the main logic of the program.

This problem might be addressed by somehow ⁴labeling each program statement by its category, or at least by whether it has an impact on data or control flow. Being able to skip irrelevant lines might make it much easier to read the important parts of a method. This information might be determined using some sort of analysis that identifies statements that ⁵do not affect the output of the method, or by some sort of machine-learned model that predicts which cross-cutting concern a given statement relates to.

Problem 3: Reading class declarations. Along the same lines as problem 2, I have found that most classes in large projects fall into a few different categories, and quickly identifying which category a class belongs to can make it much easier to read the implementation. Some categories are: data (typically immutable objects that represent something outside the program, like Dog or

Summary of Comments on Ioncaric.pdf

Page: 1

 Number: 1 Author: mernst Subject: Highlight Date: 1/11/2016 7:55:11 PM

 Number: 2 Author: mernst Subject: Highlight Date: 1/11/2016 7:58:04 PM

This is an interesting problem, though I'm not sure how you would apply AI to solve it. It seems more like a straight program analysis task.

I'm not sure whether it is relevant, but you might look at Philip Guo's work on avoiding re-computation in Python scientific programs. I don't know how much program analysis it has, actually.

Would it have been possible to only run your program on the files with a particular failure, if that's the one you had just tried to fix?

Would it have been possible to

 Number: 3 Author: mernst Subject: Highlight Date: 1/11/2016 7:58:31 PM

Right. In general, code is read much more often than it is written. If code isn't red or, that's an indication that the code was not successful.

 Number: 4 Author: mernst Subject: Highlight Date: 1/11/2016 8:00:47 PM

This sounds a bit like you want to generate aspects. I presume you would then have wanted your IDE to elide many of the aspects.

 Number: 5 Author: mernst Subject: Highlight Date: 1/11/2016 8:01:25 PM

Many methods are intended to have side effects; setters are one example, but there are others. See you might need to sharpen this criterion.

Webpage), data structures (mutable objects that organize data in some fashion), and workers (classes with relatively little state that implement a few functions for computing values). Different projects may have their own project-specific categories as well, like classes related specifically to logging or to interface with a database. This categorization is a fuzzy notion, but it might be possible to learn some features that distinguish one category from another and train a model to **1** identify the category in most cases.

Number: 1 Author: mernst Subject: Highlight Date: 1/11/2016 8:04:56 PM

This is possible. I agree that knowing which category a class is in is useful, but I wonder if it's really the most important thing. I suspect you were preaching to better understand the overall structure, and this was just one minor obstacle along the way, albeit an irritating one. Can you think of a higher level task or what your aim as a programmer was?