Lovecraft: Crowdsourced, pattern-based code improvement

> An idea for a large-scale system that identifies and suggests potential improvements for any given source code repository, based on patterns in the commit history of many other repositories.

GitHub has more than two million users, and host more than four million repositories. Some repositories are very active and others less so, but the daily total number of commits is massive. Each of these commits contains someone's idea of an improvement to the previously existing code.

Many ideas are highly specific to the project they apply to, though far from all. We can think of an opposite part of the spectrum, where ideas are entirely generic. Whether such ideas even exist is a question of philosophical nature, but we can in any case imagine a range of ideas inbetween the unique and the generic. Ideas to a lesser or greater degree applicable to projects other than the initial one.

The concept behind Lovecraft is to analyze a very large number of code commits, looking for patterns. It may be easiest to illustrate this by an example. A repository features this piece of code:

```
found = false
for item in list_of_items:
	if item.is_the_one:
		found = true
if found:
	do something
```

Someone notices the rest of the loop can be skipped when the item is found and commits updated code, adding a break statement:

```
found = false
for item in list_of_items:
	if item.is_the_one:
		found = true
		break
if found:
	do something
```

Now, if Lovecraft would look at this change from a generic point of view, disregarding variable names and such, it may turn out that a few thousand other commits in various repositories make the exact same change. And that another thousand repositories have code like the first piece, that could benefit from the one and same change.

Lovecraft connects occurences like this, but on a far more complex level. Based on knowns such as the language of the code, patterns are worked out from billions of commits and any projects that might benefit from found improvements are notified. Quantities also play a part; the more common a certain type of code change is, the more likely is it to be suggested.

In summary, Lovecraft is something similar to static code analysis, but with the world's collective programming knowledge as its brain. It could spare people some time and computers some cycles.

Someone with the means ought to build it.

Markus Amalthea Magnuson
markus.magnuson@gmail.com

Last updated
2012-12-02