Dear ***:

My name is Xin Tan, majored in Computer Science, Peking University, China. I am a first year PHD student. Recently I'm doing a research on OpenStack about the contribution composition of a code file, which in order to explore the contribution pattern of different kinds of files. Further, I want to locate the risk file to give developers some useful information.

***, I know that you are an active contributor to the nova repository. I wonder if I could show you my study, including some metrics to describe the contribution composition of a code file. I would appreciate it if you could show your opinions or give some advices, which would really, really help me a lot. And it would only take you a little time.

Thank you so much for your kindness.

First of all, I would give you a brief introduction to my study. We all know large software projects are usually composed of some general file types. For different file types, the contribution composition of them are various. For example, the files which are responsible for implementing core functionalities are usually very complex and they are modified very often, of course, the number of contributors are very high. However, relatively independent files are usually maintained by a small number of people who have high ownership of the code. We analysis the contribution composition of different file types, which in order to:

1) knowing the contribution composition of files in real time.

2) exploring the contribution pattern of different kinds of files.

3) Locating the risk file.

First, we define three metrics to describe the contribution composition of files.

1) Centrality:

The Centrality of a file refers to the proportion of ownership for the contributor with the highest proportion of ownership, which is calculated by the number of commit times.

2) Diversity:

We measured the uncertainty in a code file's contributions (or the diversity of sources of contributions) in a given period using the Teachman/Shannon entropy index, a commonly used diversity measure in many scientific disciplines.

$H(x)=E[I(xi)]=E[\log(2,1/p(xi))]=-\sum p(xi)\log(2,p(xi))(i=1,2,..n)$,

We assume that the more diverse the contribution, the more bugs the code file would have in this release. And We have proven that there is a significant positive correlation between the contribution diversity and the amount of defect of the file.

3) Stability:

The Stability of file means its personnel scheduling. It calculated by the total number of the contributors of this file who leave or join relative to the previous cycle. When the number of contributors to a file is instable, it usually means the high risk.

Then, we choose a nova release for a case study. We define several different files types according to the functionalities of code file and refer to the measurement value.

File type	Example in nova
The test file for active file	nova/tests/unit/virt/libvirt/test_driver.py
Exception handling file	nova/exception.py
Privilege management file	etc/nova/policy.json
Core interface file	nova/compute/api.py
Key function implementation file	nova/compute/manager.py
Module function implementation file	nova/conductor/manager.py
Function realization file of complex module	nova/db/sqlalchemy/models.py
Module interface file	nova/api/metadata/base.py
Module test file	nova/tests/unit/conductor/test_conductor.py
Module configuration file	nova/conf/scheduler.py
Non function implementation file	requirements.txt
i18n file	nova/locale/zh_CN/LC_MESSAGES/nova.po

And we calculate the above metrics of the nova active files (of course it is not accurate, because the contribution composition is effected by many factors not only file types.). We find three patterns.

Centrality: low Diversity: high Stability: low	Metric_1<=0.2 3=<Metric_2 14=<Metric_3	Key function implementation file Function realization file of complex module The test file for active file Exception handling file Privilege management file Core interface file
Centrality: medium Diversity: medium Stability: medium	0.2< Metric_1 <=0.7 2<= Metric_2<3 5=< Metric_3<14	Module function implementation file Module interface file Module test file Module configuration file
Centrality: high Diversity: low Stability: high	0.7< Metric_1<=1 0<= Metric_2<2 0=< Metric_3<5	Non function implementation file i18n file

For locating high risk file, I have two points.

1) Pattern 1(Centrality: low/ Diversity: high/ Stability: low) should be paid much more attention. But there are special cases, for example, Exception handling file, although it is modified too often, it is not complex, so the risk of it is low.

2) If the contribution composition of a file are significantly various between two cycle, it should be paid more attention to.

Ok, that's almost what I'm doing. I hope that I have expressed my ideas clearly. And I really hope to know what you think about my work on the following three questions, which would give me great help on my research:

1) Do you think the metrics are useful for developers and project managers in some way?

   In particular, could the Centrality be used to identify the experts of the file and how would it help in practice? And do you think that files with high code ownership would result in higher code quality and fewer failures?

   Do you think the contribution diversity could act as an indicator for high risk of lower code quality of the file in some way and why? And what would it mean in practice when the contribution diversity of a file changes a lot?

   Do you agree that when contributors left the project, their code would be hard to be maintained by others, and contributions made by newcomers would be more likely to bring bugs to the files? So would it help by knowing how many people left the project and how many people are newcomers to the projects and who are them? If yes, how would it help in practice?

2) What's your opinion of exploring contribution composition of file from different file types is reasonable?

3) Do you think the type files I divided is reasonable?

4) For different type of files, based on your developing experience, what's the idea contribution composition pattern of different file types?

5) Any other suggestion or ideals?

Again, I would appreciate it a lot if you could give me some advices. And thank you so much for your time.

Looking forward to your reply. Wish you have a good day.

Best regards!

Xin Tan
Department of Computer Science
School of Electronics Engineering & Computer Science
Peking University
Beijing 100871, China
E-mail：tanxin16@pku.edu.cn