OUKA
The University of Osaka
Institutional Knowledge Archive

Title	Large-scale Analysis of Software Reuse for Code and License Changes
Author(s)	Wu, Yuhao
Citation	大阪大学, 2019, 博士論文
Version Type	VoR
URL	https://doi.org/10.18910/72580
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

https://ir.library.osaka-u.ac.jp/

The University of Osaka

様式3

論 文 内 容 の 要 旨

氏 名 (Wu Yuhao)	
論文題名	Large-scale Analysis of Software Reuse for Code and License Changes (コードやライセンスの変更に関する大規模なソフトウェア再利用分析)

論文内容の要旨

Code reuse is a very common practice in software engineering. When performed in a correct way, code reuse can help developers create software products with higher quality more efficiently. Although code reuse is beneficial in many aspects, there are also several issues that we need to take special care of. One important aspect is software license, without which source code cannot be reused legally. Another aspect is the efficiency or barriers during the process code reuse.

In the first part of this dissertation, we deal with the issue of software license. Software license is a written text that grants the permissions of reusing and redistributing the software to its users. Removing or modifying the license statement by re-distributors will result in the inconsistency of license with its ancestor, and may potentially cause license infringement. However, in our study we have encountered cases where multiple source files that have the same source code but are under different licenses. Therefore, we describe and categorize different types of license inconsistencies and propose a method to detect them. Then we applied this method to Debian 7.5 and a collection of 10,514 Java projects on GitHub and present the license inconsistency cases found in these systems. With a manual analysis, we summarized various reasons behind these license inconsistency cases, some of which imply potential license infringement and require attention from the developers. This analysis also exposes the difficulty to discover license infringements, highlighting the usefulness of finding and maintaining source code provenance.

In the second part of this dissertation, we deal with the barriers during the process of code reuse. Although code reuse is a common practice, the process is not fully studied: how often and why is the source code changed during code reuse, what hinders code reuse and how can we improve it? In order to address these issues, we conduct an empirical study on code reuse from Stack Overflow, a question and answer (Q&A) platform for software developers. In this study, we first conduct an exploratory study on 289 files from 182 open source projects, which contain source code that has an explicit reference to a Stack Overflow post. We found that code modification during code reuse is a frequent action. Meanwhile, developers also write (re-implement) source code from scratch based on the idea from Stack Overflow. To further understand the barriers of reusing code and to obtain suggestions for improving the code reuse process on Q&A platforms, we conducted a survey with 453 open source developers who are also on Stack Overflow. We found that the top 3 barriers that make it difficult for developers to reuse code from Stack Overflow are: (1) too much code modification required to fit in their projects, (2) incomprehensive code, and (3) low code quality. We summarized and analyzed all survey responses and we identified that developers suggest improvements for future Q&A platforms along the following dimensions: code quality, information enhancement & management, data organization, license, and the human factor. Our findings can be used as a roadmap for researchers and developers to improve code reuse.

様式 7

論文審査の結果の要旨及び担当者

氏　名　（　Wu Yuhao　）		
	（職）	氏　名
論文審査担当者		
主 査	教授	井上 克郎
副 査	教授	楠本 真二
副 査	教授	伊野 文彦

論文審査の結果の要旨

本学位論文を審査担当者間で精査した結果、以下に述べる内容を確認した。

本学位論文は、プログラムのソースコードの再利用における潜在的な問題を発見し、それを解決することを目的として、コードの再利用のライセンス面とソースコード再利用方法に関する以下の2つの研究を行った。

一つ目は、本学位論文2章の、オープンソースライセンス不一致問題に関する研究である。オープンソースライセンスはコードの再利用において重要なものであり、それに違反すると開発者に大きな損害をもたらす可能性がある。従来研究では、ライセンスの検出や整合性において、いろいろなツールや手法が提案されているが、同じソースファイルに異なるライセンスが含まれてるというライセンス不一致問題はまだ不明である。そこで本研究は、ライセンス検出ツールとコードクローン検出ツールを用い、大規模なオープンソースプロジェクトにおけるライセンス不一致のソースファイルを検出する手法を提案した。実装したツールでLinuxディストリビューションのDebianを分析した結果、ライセンス不一致は多数存在し、その原因は多様であることがわかった。そしてこの問題は、開発者たちが注意すべき大きな課題であることが分かった。

二つ目は、同3章の、コードの再利用における困難さや障壁を発見し解決する方法を見出す研究である。コードの再利用は、ソフトウェア開発の効率や品質を向上させる有効な手法の一つである。従来研究では、開発者Q&Aプラットフォームを用いて開発知識やソースコードの再利用を支援する手法を提案していたが、再利用時に開発者がどういう困難に直面したか、どうすればコードの再利用をよりスムーズに進めることができるかは不明であった。そこで本章では、Stack Overflowという開発者Q&Aプラットフォームから、GitHubに公開されているオープンソースプロジェクトに利用されている事例を分析し、5つの再利用のタイプを類別した。その結果、修正が必要な再利用は３２％であり、修正の支援方法が重要であることを示した。さらに、開発者への大規模な調査を行った結果、再利用時の問題として、１）コードを自分のプロジェクトに再利用するには、必要な修正は多すぎる；２）コードが理解できない；３）コードの品質が低い、ということがあることが判明した。これらの調査で収集した情報は、次世代Q&Aプラットフォームの開発に有用である。

以上のことから、本学位論文は上記課題に対して、積極的に解決策を提案/評価し、顕著な成果を上げていることを確認した。また、ソースコードの再利用における課題を解決し、本分野へ貢献しているものであるといえる。よって、博士（情報科学）の学位論文として価値のあるものと認める。