OUKA
The University of Osaka
Institutional Knowledge Archive

Title	Study on Reuse Issues in Open Source Software
Author(s)	仇, 実
Citation	大阪大学, 2021, 博士論文
Version Type	VoR
URL	https://doi.org/10.18910/82284
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

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

The University of Osaka

論 文 内 容 の 要 旨

氏 名 (仇 実)	
論文題名	Study on Reuse Issues in Open Source Software (オープンソースソフトウェアの再利用におけるいくつかの課題に関する研究)

論文内容の要旨

Open source software (OSS) is software whose source code can be reused under some particular terms and conditions. Reusing OSS in software development has long been proved to be a good method to increase software productivity. Although OSS reuse can help developers create software products more efficiently, it also comes with some reuse issues needed to pay attention to. One important aspect is software license, without which under what conditions OSS can be redistributed, reused, and modified can not be understood correctly. Another aspect is software copyright, without which the copyright owner who owns the right to determine these conditions can not be identified clearly. Furthermore, understanding the popularity growth of OSS is also important for maintenance and reuse.

In the first part of this dissertation, we deal with the issue of software copyright. Software copyright claims an exclusive right for the software copyright owner to determine whether and under what conditions others can modify, reuse, or redistribute this software. For OSS, it is very important to identify the copyright owner who can control those activities with license compliance. In this part, as the first step of understanding copyright owner, we will explore the situation of the software copyright in the Linux kernel, a typical example of OSS, by analyzing and comparing two kinds of datasets, copyright notices in source files and source code contributors in the software repositories. The discrepancy between the two kinds of analysis results is defined as copyright inconsistency. The analysis result has indicated that copyright inconsistencies are prevalent in the Linux kernel. We have also found that code reuse, affiliation change, refactoring, support function, and others' contributions potentially have impacts on the occurrence of the copyright inconsistencies in the Linux kernel. This study exposes the difficulty in managing software copyright in OSS, highlighting the usefulness of future work to address software copyright problems.

In the second part of this dissertation, we deal with the difficulty of copyright notice identification of source files. For OSS, identifying copyright notices is important. The copyright owner is allowed to change its license or to grant a commercial one to a third party and start legal proceedings to enforce its license. Several licenses also require that the copyright owner of OSS projects being reused should be recorded explicitly. However, both the collaborative manner of OSS project development and the large number of source files increase its difficulty. In this part, we aim at automatically identifying the copyright notices in source files based on machine learning techniques. The evaluation experiment shows that our method outperforms FOSSology, the only existing method based on the regular expression.

In the third part of this dissertation, we deal with the issue of software license. Software license is a written text that claims under what conditions others can modify, reuse, or redistribute the software. A license may violate another one according to the terms and conditions. Making software by reusing OSS as dependency may cause dependency-related license violation if the developers overlook the license of the dependency. In this part, we first conduct an empirical study on npm - a JavaScript-based software ecosystem - to study the prevalence of dependency-related license violations. The result suggests that only a few packages (0.644%) in npm have dependency-related license violations. However, we also observe that including the packages licensed under copyleft licenses in the dependency network highly potentially causes dependency-related license violations. We then conduct a preliminary questionnaire on the authors of packages detected as having dependency-related license violations to study the developers' attitudes. The results reveal: 1) the developers' overlooking and misunderstanding of the dependency-related license violations; 2) the difficulties in managing dependency-related license violations and the developers' demands for help.

In the fourth part of this dissertation, we deal with the issue of software popularity. In OSS ecosystems, software popularity is valuable information to developers because they continually want to know whether their software is attracting and gaining acceptance. Software popularity is also an important indicator to suggest if a software is beating its competitors in an OSS ecosystem. Meanwhile, the software with the rapid growth of popularity is a double-edged sword for OSS ecosystem and its reusers. Accordingly, it is important to understand the popularity growth of packages (i.e., how fast packages become popular). In this part, we conduct an exploratory study on packages in the node package manager (npm) to understand: (1) the characteristics of popularity growth, and (2) the factors that could affect popularity growth. We propose a method to model popularity growth as a curve

論 文 内 容 の 要 旨

氏 名 （ 仇 実 ）	
論文題名	Study on Reuse Issues in Open Source Software （オープンソースソフトウェアの再利用におけるいくつかの課題に関する研究）

論文内容の要旨

and find that popularity growth mathematically follows three models – accelerated growth model (i.e., quadratic model), steady growth model (i.e., linear model), and decelerated growth model (i.e., square root model). The results show that 51.56% of the studied packages depict steady growth model, followed by accelerated growth model and decelerated growth model, 40.02% and 7.20% respectively. Furthermore, we reveal that factors including age, dependents, new features, and functionalities have impacts on popularity growth. Our study shows potential tips for helping practitioners in developing and evolving packages in a competitive OSS ecosystem.

The findings in this dissertation will help practitioners who reuse OSS in practice and researchers who are to create a better platform for OSS reuse.

様式７

論文審査の結果の要旨及び担当者

氏　名　（　仇　実　）		
	（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、オープンソースソフトウェア（OSS)の再利用時に関する課題に関して以下の３つの研究を行った。

1つ目は、学位論文の2章で示すように、OSSの著作権に関する研究を行った。OSS の代表的な例である Linux カーネルのソースコードを分析し、ソースファイル上の著作権者とリポジトリ上のソースコード貢献者の 2 種類のデータを比較した。これらが不一致の場合を著作権の不整合と定義し、Linuxカーネルでは著作権の不整合が蔓延していることが明らかになった。また、コードの再利用、所属変更、リファクタリングなどが、Linuxカーネルにおける著作権の不整合の発生の原因であることを明らかにした。本研究は、OSSにおけるソフトウェアの著作権管理の難しさを示した。

2つ目は、3章で示すように、ソースファイル上の著作権表示の識別の研究を行った。OSSでは、著作権表示の識別が重要である。著作権者は、そのライセンスを変更したり、商用のものを第三者に付与したりして、そのライセンスを強制するための法的手続きを開始することができる。また、いくつかのライセンスでは、再利用されるOSSプロジェクトの著作権者を明示的に記録する必要がある。しかし、OSSプロジェクトでは、数多くの人が開発に関わり、また、ソースコードファイルの数も膨大であり正確に著作権者を特定することは容易ではない。3章では、機械学習を用いて、ソースファイル中の著作権表示を自動的に分析する方法を提案した。評価実験の結果、本手法による著作権者の認識は、正規表現に基づく唯一の既存手法であるFOSSologyを上回ることが示された。

3つ目は、4章で示すように、ソフトウェアの再利用性の評価に関する研究を行った。OSSの評価指標としてユーザのダウンロード数（DL数）は重要な指標である。この章では、(1)DL数の上昇の特徴、(2)DL数の上昇に影響を与える要因を理解するために、ノードパッケージマネージャ(npm)のパッケージを対象とした分析を行った。その結果、DL数の成長を曲線でモデル化する方法を提案し、DL数は加速成長モデル(2次モデル)、定常成長モデル(線形モデル)、減速成長モデル(平方根モデル)の3つのモデルに従うことを示した。これらのうち定常成長モデルが51.56%を占め、次いで加速成長モデルが40.02%、減速成長モデルが7.20%となった。さらに、経過年数、依存関係、新機能、機能性などの要因がDL数の成長に影響を与えていることを明らかにした。

本論文で得られた知見は、OSSの再利用を実践する実務者や、OSS再利用のためのより良いプラットフォームを構築しようとしている研究者に非常に役立つものである。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。