

Same File, Different Changes: The Potential of Meta-Maintenance on Github

H. Hata, R. G. Kula, T. Ishio, C. Treude,

IEEE/ACM 43rd International Conference on
Software Engineering(ICSE), 2021, pp. 773-784

発表者: 井町 瑠貴

背景

GitHubなどのプラットフォームにより、開発者はリポジトリ間でコードを容易に再利用・共有できるようになった

これらの共有ファイルを維持することは、特に再利用されるコードの最新版を維持するために重要である

背景

同じ起源を持つ様々なリポジトリから有用な変更点を集約して構成することができれば、複数のプロジェクトのソフトウェア保守活動を効率的に行うことができる

このアプローチをメタ・メンテナンスと呼ぶ

目的

GitHubにおける現代のソフトウェアプロジェクトのためのメタ・メンテナンスの可能性を探る

メタ・メンテナンスに関する用語

- ・シードファイル

→複数のリポジトリで共有されている特定のバージョンのファイルのこと
同じシードファイルを含むリポジトリを、
同じシードファミリーに属していると呼ぶ

- ・バリエーション

→同じシードファイルからの更新ファイルのこと

リサーチクエスト

RQ1: ソフトウェアリポジトリにおける
シードファイルの普及率はどの程度か?

RQ2: シードファイルにはどのような種類があるか?

RQ3: シードファイルのリポジトリ間の
関係はあるか?

リサーチクエスト

RQ4: シードファイルの変化の主な要因は何か?

RQ5: シードファミリーの中で、バリエーションはどのようにして独自に進化するか?

RQ6: 開発者はバリエーションの変更をどのように考えているか?

データセット

- 対象は, C, C++, Java, JavaScript, Python, PHP, Rubyの7つの言語で書かれたGitHub上のソフトウェア開発リポジトリ
- GHTorrentのデータセットを用いて,
以下の基準でアクティブなリポジトリを特定
 1. 履歴全体で500回以上のコミットがある
 2. 最も活発な2年間で100回以上のコミットがある

データセット

- `git rev-list -all -objects` と `git cat-file` コマンドを用いて、既存のBlobを抽出
- 複数のリポジトリに現れたBlobをシードファイルとみなす
- 合計で27,994,587個のシードファイルが得られた

データセット

- ・シードファイルを3つに分類
→common, sometimes, rare

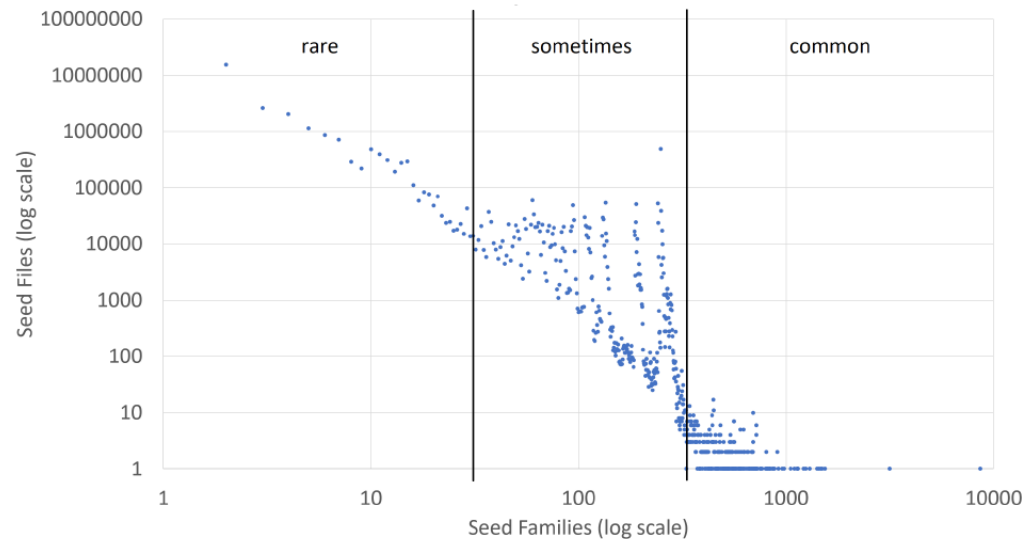


Fig. 1: Plot of the number of seed families for each number of seed files (log-log scale).

アプローチ

- ・定量的分析(RQ1, RQ5)

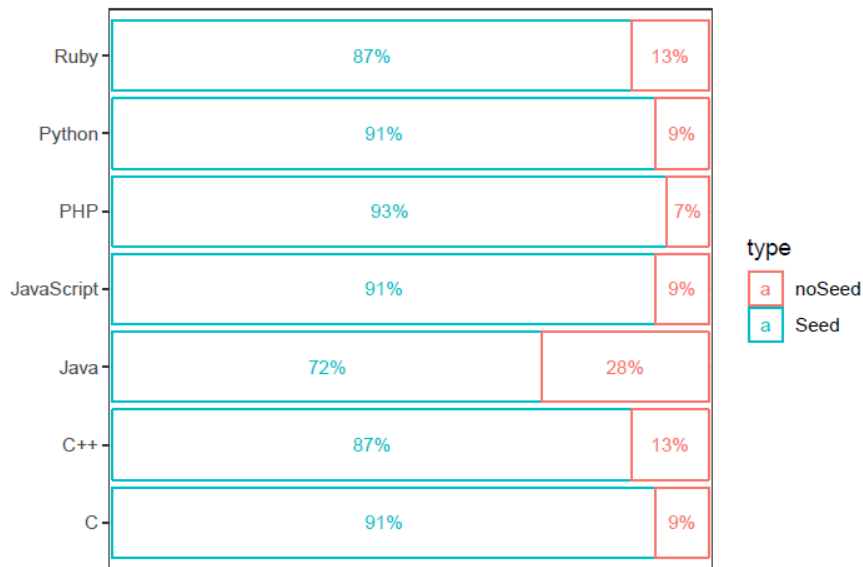
収集したデータセットとサンプルをシードファイルの存在, 状態, サイズ, 変更の観点から層別して分析を行う

また, 「シードファイルとの類似性の保持」と「他のバリエーションとの違いの独自性」という指標を用いて, 各ファミリーにおけるシードファイルの進化の度合いを測定

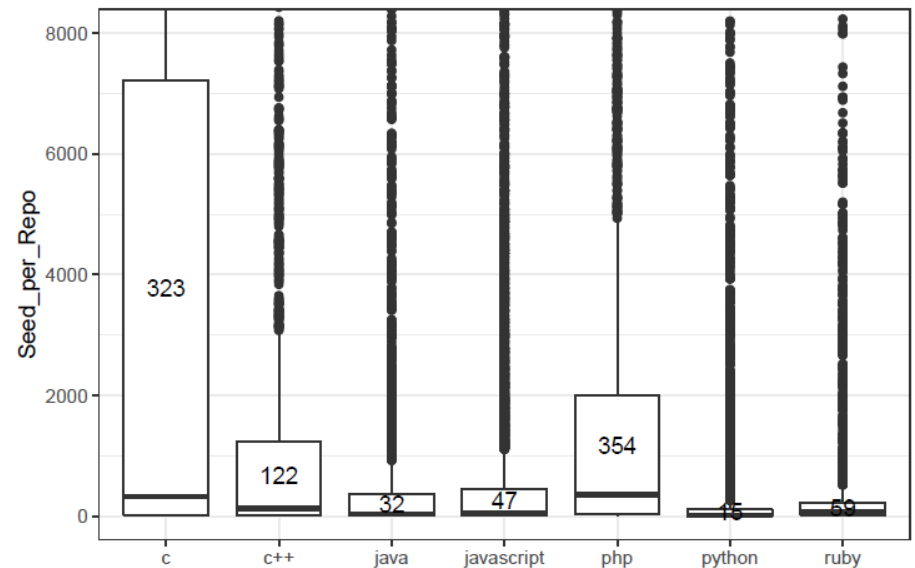
アプローチ

- 定性的分析 (RQ2, RQ3, RQ4)
3人の著者で手動で複数回アノテーションを行う
- アンケート調査 (RQ6)
メールなどでメタメンテナンスに関する
アンケートを配布
→3件の回答が得られた

RQ1: シードファイルの普及率はどの程度か?



(a) Percentage of repositories with and without seed files.



(b) Distribution of the number of seed files per repository.

Fig. 2: Prevalence of seeds from (a) seed file existence and (b) the number of seed files in a repository, per language.

RQ1:

シードファイルの普及率はどの程度か?

対象となったリポジトリのうち、70%以上のリポジトリに、少なくとも1つのシードファイルが存在していた

また、大量のシードファイルが存在するにもかかわらず、そのほとんどが現在はメンテナンスされておらず使用されていない

RQ2: シードファイルにはどのような種類があるか?

TABLE VI: Types of Seed Files

	library	configuration	utility	other
common	91	5	2	3
sometimes	5	1	157	10
rare	1	1	12	2

RQ2: シードファイルにはどのような種類があるか?

- ・大規模なシードファミリー
→シードファイルは、ライブラリであることが多い
- ・中規模なシードファミリー
→ユーティリティー機能を含む傾向がある
(動作制御のためのソフトウェア)

大規模なシードファミリーのファイルは、小規模なシードファミリーのファイルに比べて、コード行数が少ない傾向にある

RQ3: シードファイルのリポジトリ間の関係はあるか?

- ・大規模なものは、関連性がない傾向にある
- ・小規模なものは、関連性がある傾向にある

TABLE VII: Relationships among Seed Families

	related	non-related
common	0	398
sometimes	624	24
rare	40	5

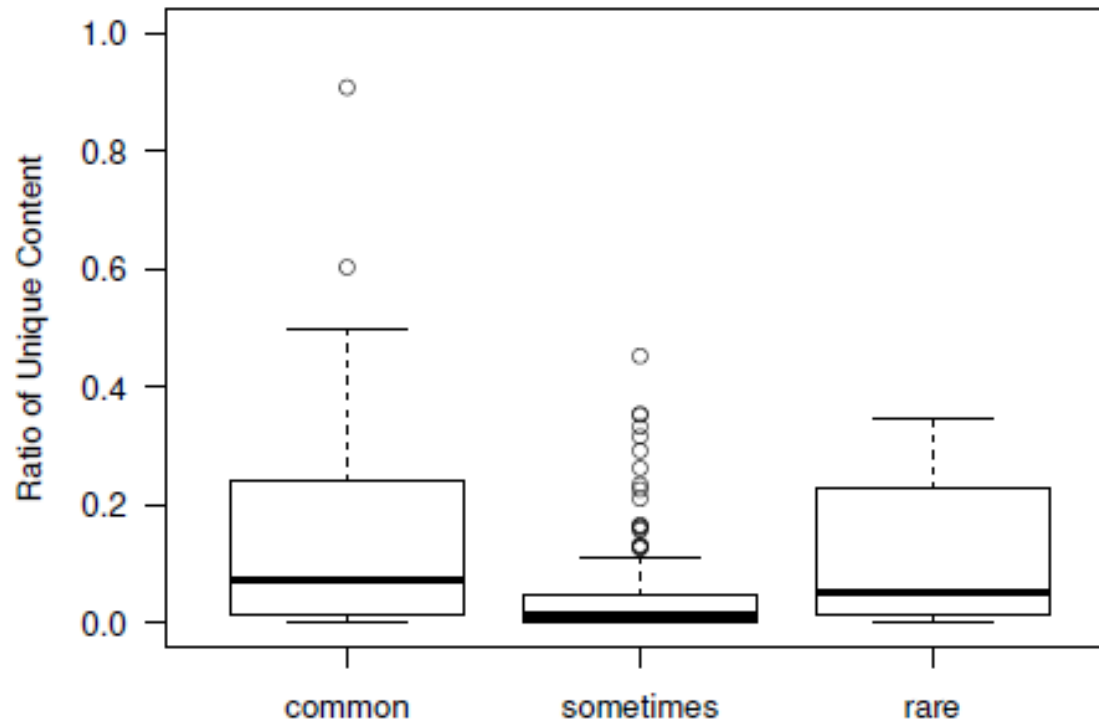
RQ4: シードファイルの変化の 主要因は何か?

大規模(common) または小規模(rare) なシードファミリのシードファイルには, メタ・メンテナンスに有用なプロジェクト固有の変更も多く含まれていた

TABLE VIII: Change Types

	known origin	library updates	project- specific	tangled updates	other
common	9	202	96	15	76
sometimes	619	11	11	3	4
rare	34	0	8	0	3

RQ5: バリエントはどのようにして 独自に進化するのか?



(b) Distribution of $Uniqueness(V, s)$, per stratum.

RQ5: バリエントはどのようにして 独自に進化するのか?

大規模(common) または小規模(rare) なシードファミリーの変化は, プロジェクト固有の変更を含んでいるが, 中規模(sometimes)の変化は共通の変更を含む傾向がある

シードファイルを進化とメンテナンスの観点から理解するためには, さらなる研究が必要である

RQ6: 開発者はバリエーションの変更をどのように考えているか?

開発者からは、セキュリティやパフォーマンスなどに関する変更をサポートすることが、メタ・メンテナンスの有望なユースケースであり、開発者が望んでいることであることがわかった

また、リポジトリ間の関係の性質をさらに理解するための今後の課題も指摘されている

メタ・メンテナンスの課題

- ・ シードファイルのセットにおけるclone-and-ownの
関係の詳細な調査
- ・ ソースコード追跡システムを開発し, リポジトリの
関係やシードファイルの履歴を把握
- ・ バリエーション間の有用な変更点を特定する技術
- ・ メタ・メンテナンスのためのツールサポート

結論

メタ・メンテナンスの可能性を探るため、
次のような調査を行った

1. シードファイルを定量的に分析し、普及率、
進化の程度、独自性を明らかにした
2. 質的分析を行い、種類、関係、および
バリエーションの変化の主要因を明らかにした
3. 開発者からのフィードバックを得るための
アンケート調査を行った

単純なフォーク以外にも様々な種類の変更が
確認され、メタ・メンテナンスの可能性が示された

結論

また、以下のような課題がある

1. シードファミリー内のすべてのシードバリエーションを管理する方法の理解
2. 有用な変更とは何かについてのさらなる研究
3. シードファミリーのニーズを抽出して他のリポジトリに照会するためのツールサポート

所感

- アンケート調査の回答数が少ないため、まだ他にも課題や開発者の望んでいることがあるかもしれないと感じた
- GitHub以外のサービスや、他の言語を対象とするとどうなるのか気になった