

How Often and What StackOverflow Posts Do Developers Reference in Their GitHub Projects?

S. S. Manes and O. Baysal

2019 IEEE/ACM 16th International Conference on Mining
Software Repositories (MSR), 2019, pp. 235-239

発表者: 井町 瑠貴

背景

Stack Overflow (SO) は、開発者が直面する問題の解決策を見つけるための Q&A フォーラムである

投稿は、コードスニペット、テキストの説明、さらに参照するための外部ソースへのリンクで構成

開発者は、SO のコードスニペットを、保守性やライセンスを気にすることなく使用、適応できる

背景

GitHubプロジェクトでは, SOの投稿を約650万件参照

OSSプロジェクトとSOの議論が相互に依存していることから, GitHubのコードリポジトリにおけるSOの参照の性質を調査する実証的な研究を行う

リサーチクエスト

RQ1: GitHubの開発者は、コードの中でどのくらいの頻度でSOの投稿を参照しているか?

RQ2: SOを参照しているGitHubプロジェクトの特徴はなにか?

RQ3: GitHubプロジェクトでは、どのような種類のSOの議論が最も人気があるか?

RQ4: コードスニペットを使ったSOの議論は、時間とともに進化するか?

アプローチ

- SOTorrent: SOの公式データダンプに基づいたオープンデータセット
- GHTorrent: GitHubのクエリ可能なオフラインミラー
- これらをマッピングして分析を行う

RQ1:

GitHubの開発者はどのくらいの頻度でSOの投稿を参照しているか?

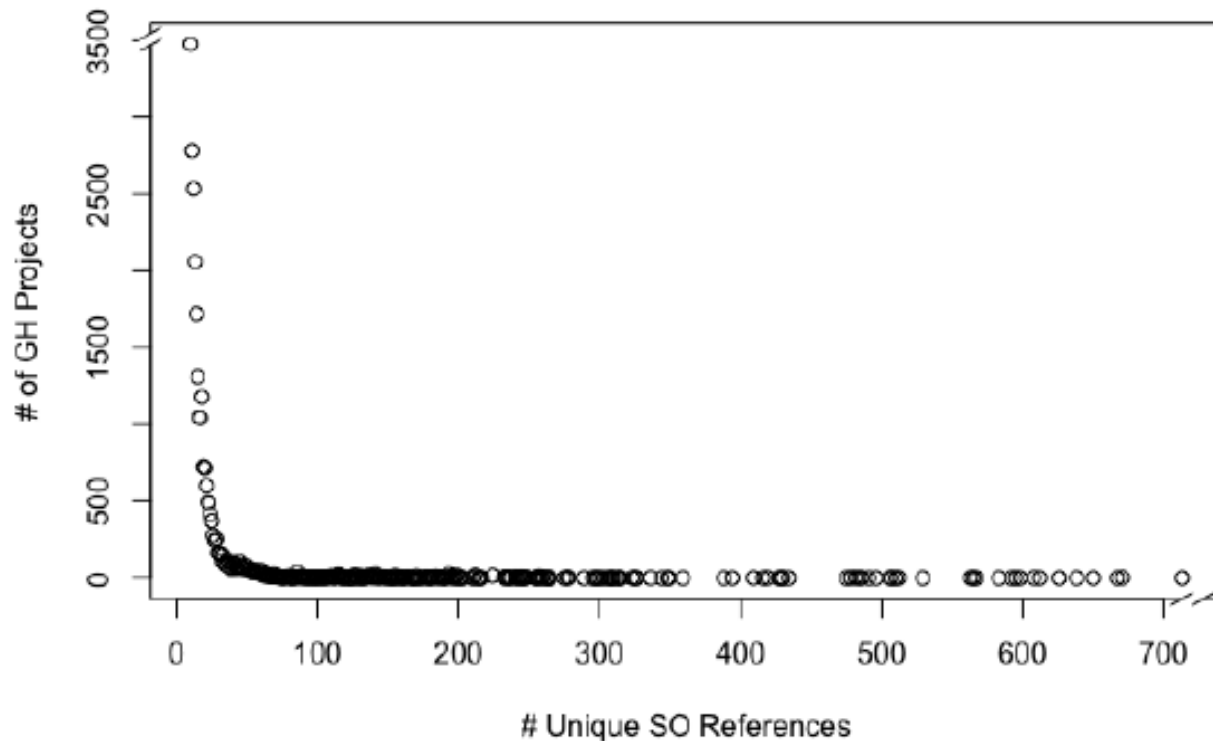


Fig. 1: Number of SO references in GitHub projects.

RQ1:

GitHubの開発者はどのくらいの頻度でSOの投稿を参照しているか?

- ・ほとんどのプロジェクトでは、開発者1人あたり約30回参照しているが、一部の開発者はコード内で250回以上SOの議論を参照している
- ・プロジェクトごとのSOの参照数の平均は176件

RQ2: SOを参照しているGitHubプロジェクトの 特徴はなにか?

1. チームサイズ

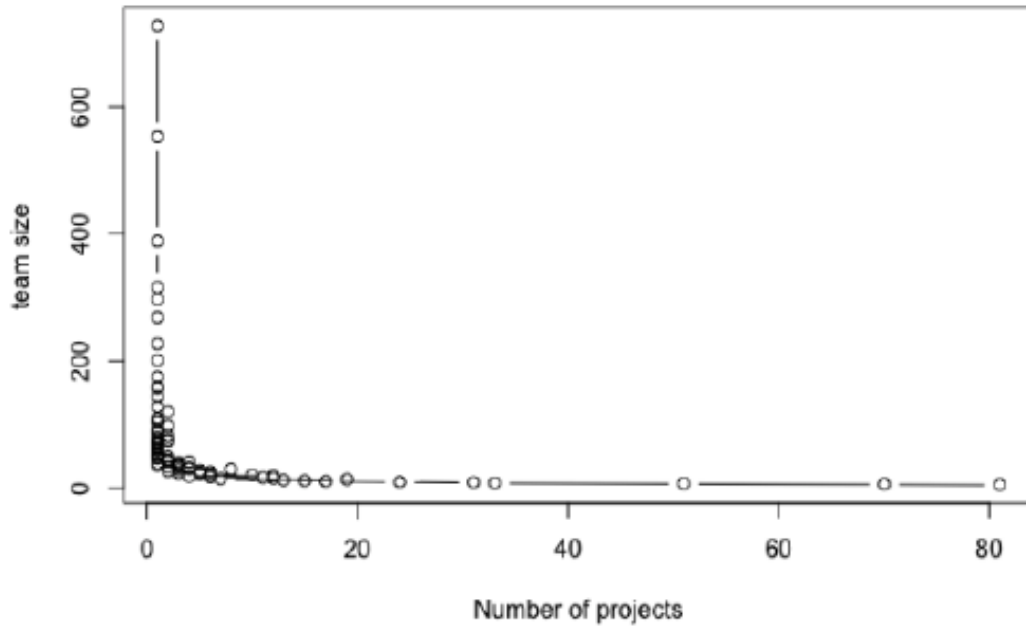


Fig. 2: GitHub team size distribution.

RQ2: SOを参照しているGitHubプロジェクトの 特徴はなにか?

1. チームサイズ

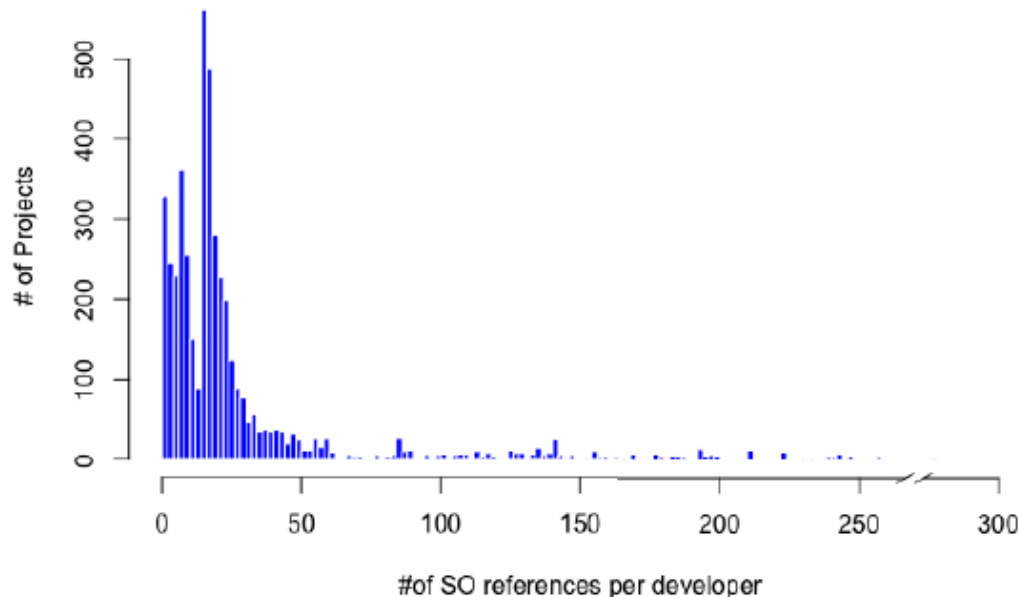


Fig. 3: How often do developers refer to SO? [Note: the # of SO references are normalized by team size].

RQ2: SOを参照しているGitHubプロジェクトの 特徴はなにか?

1. チームサイズ

- ・チームサイズは, 5人から726人の開発者の間で変化している
- ・チームの規模が大きくなると, より多くの開発者が参照されたSOの投稿について議論する



SOから借用したコードスニペットやコンセプトの
安定性にも影響を与えると推測される

RQ2: SOを参照しているGitHubプロジェクトの 特徴はなにか?

2.プログラミング言語

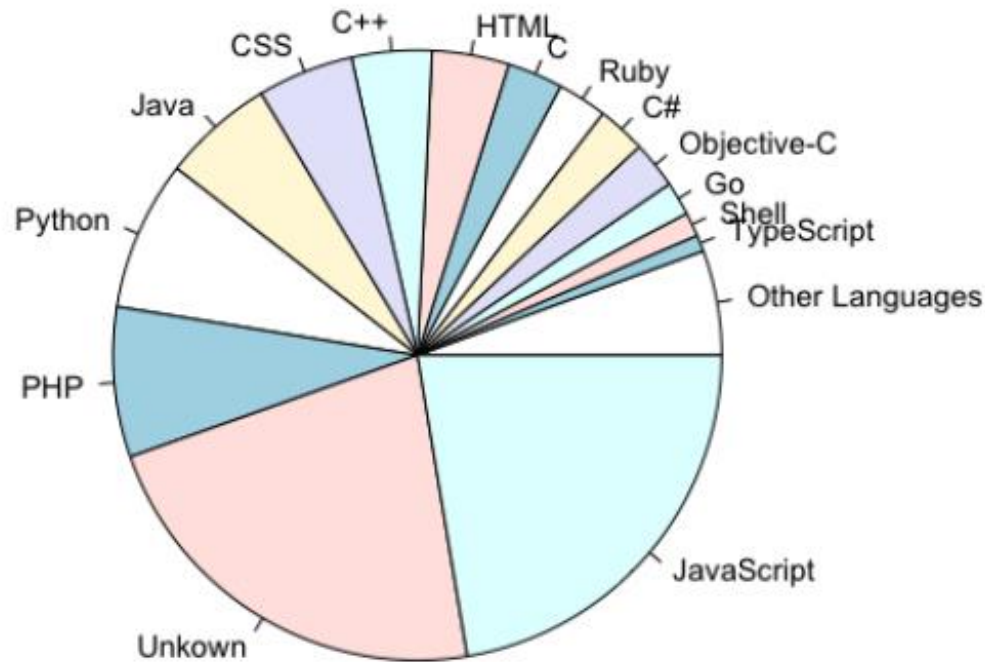


Fig. 4: Language distribution in GitHub projects.

RQ2:
SOを参照しているGitHubプロジェクトの
特徴はなにか?

2.プログラミング言語

- ・結果は, GHTorrentのProjectsテーブルの言語フィールドにマッピングされる
- ・SOの議論を参照しているGitHubのコードリポジトリの大部分は, JavaScript, PHP, Python, Java, CSS, C++, HTML, C, Ruby, C#などの言語に関連している

RQ3:

GitHubプロジェクトでは, どのような種類のSOの議論が最も人気があるか?

- ・言語タグとコンセプトタグに分類したSO投稿のタグに基づいて分析
- ・コード開発に関連するものののみを分析したいため, プラットフォーム, フレームワーク, ツールに関連するタグを手動で除外

RQ3:
GitHubプロジェクトでは, どのような種類の
SOの議論が最も人気があるか?

1. 言語タグ

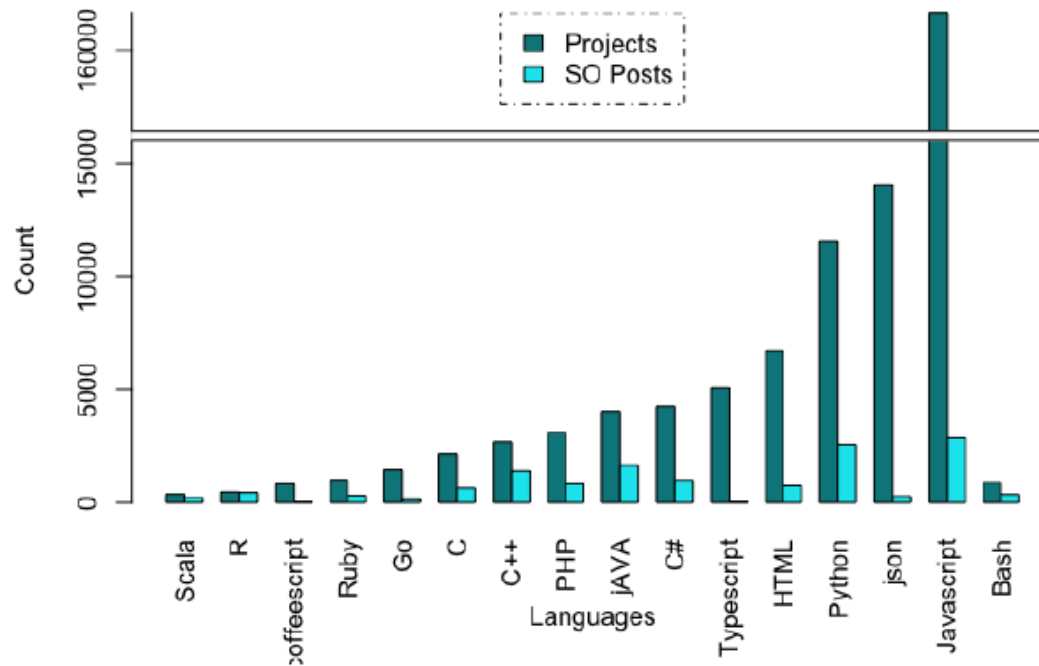


Fig. 5: Distribution of language tags in SO posts and GitHub projects.

RQ3:

GitHubプロジェクトでは, どのような種類のSOの議論が最も人気があるか?

1. 言語タグ

- ・データセットに含まれるSO投稿の79.5%は, プロジェクトの言語と同じ言語タグを持っている
- ・残りの20.5%のSO文献は, コードスニペットが別の言語ドメインに適応された場合に分類
- ・プロジェクトの言語分布は, typescriptとjsonを除いて, SOの投稿の言語タグ分布と類似している

RQ3:
GitHubプロジェクトでは, どのような種類の
SOの議論が最も人気があるか?

2.コンセプト

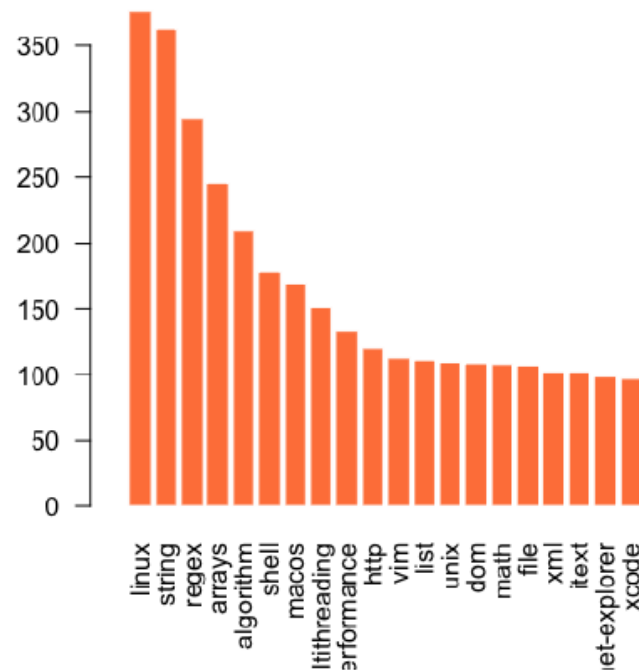


Fig. 6: Most popular SO concepts being borrowed by GitHub developers.

RQ4:

コードスニペットを使ったSOの議論は、
時間とともに進化するか？

- ・投稿のコードブロックのバージョンを分析することで追跡可能なその進化について考える
- ・コードスニペットの進化を判断するため、受け入れられた回答を考慮
- ・そのため、回答投稿におけるコードスニペットのバージョンに注目

RQ4:
コードスニペットを使ったSOの議論は、
時間とともに進化するか？

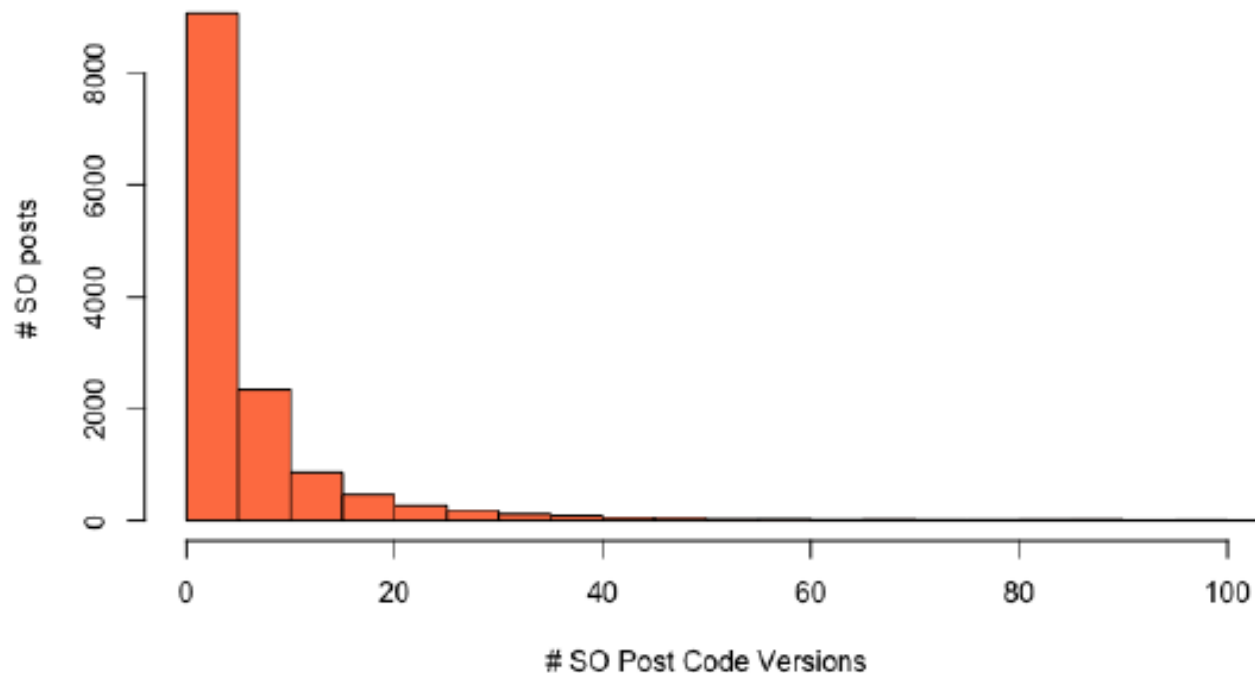


Fig. 7: Evolution of code snippets in SO posts.

RQ4:

コードスニペットを使ったSOの議論は、
時間とともに進化するか？

- ・全投稿のうち、コードスニペットが少なくとも1回変更されたのは79%、20回以上変更されたのは9%であった
- ・SOからコードスニペットを採用する場合、SOでのコードスニペットや議論が時間とともに変化していく可能性があることに注意する必要がある

注意点

- ・GHTorrentにプロジェクト貢献者の最新データがない
- ・各プロジェクトはその出自を考慮せずに独立して検討・分析された
 - フォークされたプロジェクトと重複したSOの参照がある可能性がある

まとめ

- SOTorrentとGHTorrentをマッピングすることで、GitHubの開発者がSOの投稿をどのように参照しているかを研究
- コードスニペットを含む投稿の79%は、時間の経過とともに進化していることがわかった
- 今後は、SOでのコードスニペットの進化について理解を深め、プロジェクトの開発やメンテナンスにどのような影響を与えるかについて研究していく

所感

- ・自分が想像していたよりも多くのSO参照が行われていた
- ・投稿が時間の経過とともにどのように変化していくのか気になった
→投稿されてから時間が経てば経つほど変更は起きにくくなると考えられる