OSSにおける開発知識の遍在に関する実証分析

西中 隆志郎[1,a)]　山下 一寛[1,b)]　鵜林 尚靖[1,c)]　亀井 靖高[1,d)]

概要：ソフトウェアの開発履歴は，新規にソースコードを作成，修正する開発者にとって有益である．しかし開発履歴はソースコードのコミットごとの差分の蓄積であるため，バグ修正箇所以外の差分も含まれており，バグ修正箇所のみを探すのには時間がかかる．そこでバグ混入状況と修正状況の情報を含む Q&A サイトに注目する．本研究では OSS リポジトリのバグ修正データを Q&A 形式の記事データに変換することで，開発者にとって有用な開発知識を取り出す．またリポジトリのバグ修正データはすべてが開発知識として有用である訳ではないため，変換した Q&A 形式データのうち開発知識として有用なものの数を実証的に分析し，OSS の規模における開発知識の遍在の仕方を確かめる．

1. はじめに

開発者がバグ修正を行う過程で得る知識はコーディングの技術を向上させ，新たなバグを瞬時に修正し，また未然に防ぐことが可能となる．そのためソースコードの記述と修正の蓄積である OSS の開発履歴から取り出したバグ修正データは開発者にとっての開発知識となりうる．しかし OSS の開発履歴が持つ情報は，通常コミットごとのファイルの差分の情報である．バグ修正データはこの差分の情報の中に遍在しており，差分の情報をそのままデバッグの参考とするには余分な情報が多く，効率的な手段ではない．そこで，多くのプログラマが利用する StackOverflow[*1]，Teratail[*2] などのプログラミング情報を蓄積した Q&A サイトに着目する．StackOverflow，Teratail の登録ユーザ数は年々増加の一途を辿っており，この事実は Q&A サイトの利便性を実証している．また先行研究において Chen ら [1] は，StackOverflow の記事上のコード断片を分析し，既存のプロジェクトからバグと思われる箇所を発見する手法を提案している．

そのため本研究では Q&A 形式のデータの有用性に着目し，OSS リポジトリの開発履歴におけるバグ修正時のコード差分から取り出したバグ修正データを Q&A 形式の記事データに変換する．また，変換したデータのうちどれだけの数が開発知識として有用であるかを実証的に分析し，OSS の規模における開発知識の遍在の仕方を確かめる．

[1] 九州大学
Kyushu University
[a)] nishinaka@posl.ait.kyushu-u.ac.jp
[b)] yamashita@posl.ait.kyushu-u.ac.jp
[c)] ubayashi@ait.kyushu-u.ac.jp
[d)] kamei@ait.kyushu-u.ac.jp
[*1] http://stackoverflow.com/
[*2] https://teratail.com/

2. OSS 開発知識の抽出手法

2.1 使用するバグ修正データの種類

本研究では，生成する Q&A 形式のデータには API に関するバグ修正データを用いる方針をとる．Zhong ら [2] は，ソースファイルの半分のバグ修正に際して少なくとも 1 回 API に関する修正が行われる，と述べており，API に関するバグ修正データの重要性を裏付けている．

2.2 OSS リポジトリからのバグ修正データの取得

Q&A 形式のデータは以下の 2 段階の手順により生成する．概略図を**図 1** に示す．

(1) OSS リポジトリからバグ修正データを取得
(2) Q&A 形式データの生成

バグ修正データの取得には SZZ アルゴリズム [3] を使用する．このアルゴリズムはバージョン管理ツールに蓄積された開発履歴からバグ修正の行われたコミットとバグの混入したコミットを特定する．バグ修正前コードはバグを含む誤った記述でなければならないが，その記述が行われるコミットはバグ修正コミットの直前のコミットであるとは限らない．そのため SZZ アルゴリズムによりバグ混入コミットを探知する．得られたバグ混入コミットとバグ修正コミットのコード差分からバグ修正データとしてバグ修正前コードとバグ修正後コードが得られる．

2.3 Q&A 形式データの生成

生成する Q&A 形式データは質問 (Question) コードと回

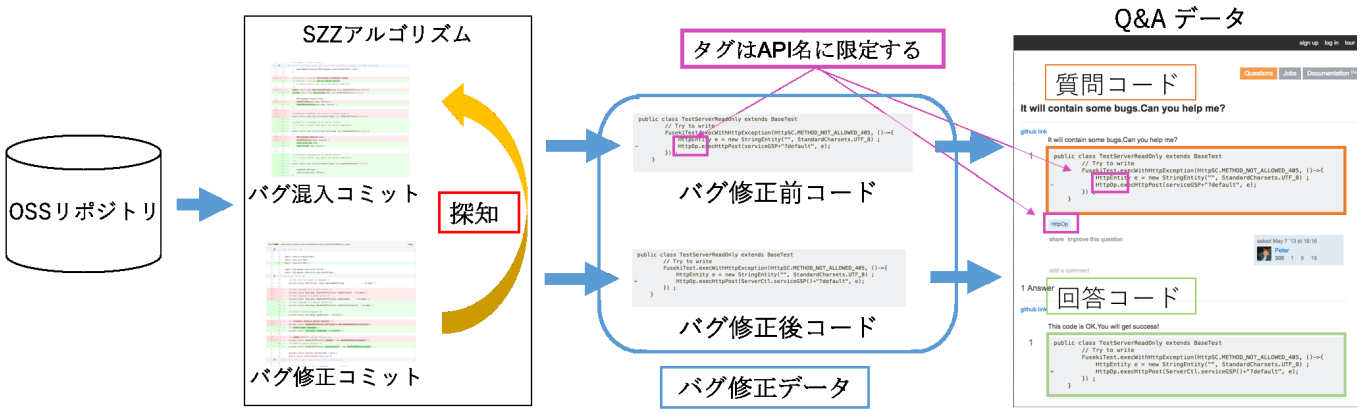

図 1 Q&A 形式データ生成の概略図

答 (Answer) コードにより構成されており，2 つのコードには OSS リポジトリから取得したバグ修正前コードとバグ修正後コードをそれぞれ割り当てる．

3. 評価方法

3.1 Q&A 形式データの評価方法

生成した Q&A 形式データを以下の 2 つの観点で分析する．

(1) 実際のバグ修正に役立つ Q&A 形式のデータがどれだけ生成できるか

(2) 実際のユーザが Q&A 形式のデータを便利に感じるか

観点 (1) のための評価方法として，バグ修正前コード同士の比較を行う．比較に用いるため，バグ修正データのバグ修正コードから API 名のタグを抽出する．テストデータのバグ修正データと Q&A 形式データ生成に用いたバグ修正データのバグ修正前コード同士を比較し，タグが一致し，かつコードクローンが存在した場合に Q&A 形式データは実際のバグ修正に役立つと定義し，そのデータの数を計測する．

観点 (2) のための評価方法として，生成した Q&A 形式のデータを実際に StackOverflow に投稿し，StackOverflow 上の記事評価システム [*3] を活用して一般ユーザーの評価を得る．

4. 現状と今後の予定

生成される Q&A 形式データの中にはライブラリの導入部分のみ変更されているデータなど，バグ修正に有用となり得ないデータがある．そこで 3.1 節の観点 (1) の初歩的な評価を行うため，Q&A 形式データを明らかに有用となり得ないものとなり得るものとに分類した．明らかに有用となり得ないデータは以下の 4 つの条件を満たすものとした．

(1) 修正前，修正後のペアになっていないデータ

[*3] http://stackoverflow.com/help/privileges/vote-up

表 1 予備実験データセット Apache Jena の情報

開発期間	NOR(NOB)	バグ混入・修正データ
2012-5-18〜2016-10-19	5,381(3,278)	34,580

NOR: Number of revisions NOB: Number of bug fix revisions

(2) 修正箇所がライブラリの導入部分のみであるデータ

(3) 修正箇所がコメント行のみであるデータ

(4) 開発者のメモなどのソースコードでないデータ

これにより有用となり得るデータに関してさらに詳細に分析を行うことができる．対象となるデータセットは Apache Jena である．データセットに関する情報を**表 1** に示す．分析により有用となり得ないデータが 36%，有用となり得るデータが残りの 64%という結果が得られ，半数を超える Q&A 形式データが有用である可能性を持つことがわかった．

研究の最終目標は OSS リポジトリから生成した Q&A 形式データを StackOverflow に投稿して Q&A 情報サイトを増強し，多くのプログラマに利益をもたらすことである．今後の研究では，生成された Q&A 形式のデータのうち有益なものの数について分析を行う．

謝辞 本研究は，文部科学省科学研究補助費 基盤研究 (A)(課題番号 26240007) による助成を受けた．

参考文献

[1] Fuxiang Chen and Sunghun Kim, Crowd Debugging, *In Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering (ESEC/FSE 2015)*, pp.320-332

[2] Hao Zhong and Zhendong Su, An Empirical Study on Real Bug Fixes, *In Proceedings of the 37th International Conference on Software Engineering - Volume 1 (ICSE '15)*, pp.913-923

[3] Jacek S liwerski, Thomas Zimmermann and Andreas Zeller, When Do Changes Induce Fixes?, *In Proceedings of the 2005 international workshop on Mining software repositories (MSR '05)*, pp.1-5