6N-08

Stack Overflow記事の充実に向けたソースコード変更履歴の利用可能性評価

西中隆志郎† 佐藤亮介† 亀井靖高† 鵜林尚靖†
†九州大学

1 はじめに

Q&Aサイトは，作業中の開発者が知識を獲得するために利用されており，特にStack Overflow（SO）[1]は，開発者に有用な様々な情報を与えるものとして期待されている[1][7]．SOはプログラミング技術に特化したQ&Aサイトであり，コード片を質問・回答投稿に含めて共有できる．例えばSO投稿の中には特定のAPIの使用法に関する投稿記事が存在する．公開されているAPIの中には，公式ドキュメントの内容が不足しているものも多く，そういったドキュメントの代替としてSOの投稿が利用されている[5]．

しかしながら，SOが得意とするプログラミング技術の領域は一部に限られている．SO投稿の量は投稿の専門分野によってばらついており，特定の分野においては投稿が存在しない場合もある[5]．また，質問投稿に回答が投稿されるまでの時間が，質問投稿の専門分野によりばらつく問題もある[3]．

そこで本研究では，SOを充実させることにより開発者を支援することを目指す．その第一歩として，本稿では，ソースコード差分の内容がSOを充実させる可能性があるか否かを調査する．ソースコード差分に着目する理由は，情報の乏しい状況下でも，試行錯誤しながら問題に対処する開発者は少なからず存在し，彼らの作業の情報はソフトウェアの開発履歴の中に存在し，抽出できる可能性があると考えたためである．ソースコード差分を用いる動機となる事例は文献[8]に存在する．

本稿では，関連性の観点としてAPIの種類の粒度に着目する．SOでは，プログラミング言語やフレームワークといったAPIの粒度でしばしば議論が行われている[6]．

Evaluation of availability of source code change histories towards enrichment of Stack Overflow contents

Ryujiro NISHINAKA†, Ryosuke SATO†, Yasutaka KAMEI†, and Naoyasu UBAYASHI†
†Kyushu University, 819-0395, Fukuoka, Japan
nishinaka@posl.ait.kyushu-u.ac.jp
{sato, kamei, ubayashi}@ait.kyushu-u.ac.jp

[1]http://stackoverflow.com/

表 1: データセットに用いるSO投稿

投稿数（質問投稿）	投稿期間	条件
51,003	2008/08/07 〜 2017/03/13	「android」および「java」タグを質問投稿が所持

2 調査と結果

2.1 データセット

ソースコード差分を生成するため，F-Droid[2]で公開されているAndroidアプリケーションリポジトリ713件を用いる．Android SDKのドキュメントは一般的に不足していると指摘されており[4]，開発者が試行錯誤しながら開発を行なう可能性があると考えたためである．また，調査対象とするSO投稿データとして，Stack Exchange Data Dump[2]で公開されている2017年3月版のダンプデータを使用する．データの詳細を**表 1**に示す．表1中の質問投稿に対応する回答投稿も調査の対象である．

2.2 調査課題とアプローチ

差分の内容がSO投稿を充実する可能性を調査するため，2つの調査課題（RQ）に答える．

2.2.1 （RQ1）ソースコード差分はどれだけの実際のSO記事に関連するのか

動機: 差分と内容が近い実際のSO記事の量は，差分がSO記事に対して情報をもたらす可能性と考えられる．ソースコード差分がSO記事を量的に支援できるかを第一に調べるため，差分と内容が近い実際のSO記事の量を計測する．

アプローチ： 内容の近さは，扱われたSDKメソッド名の列の一致で近似的に判断する．APIの粒度で関連性を計測するため，コード差分において使用されたAndroid SDKメソッドを，正規表現を用いて計測する．差分の中で使用されていたAndroid SDKメソッドのクラス名およびメソッド名が，投稿本文の文字列中に含まれるSO記事を，差分と内容の近いSO記事と判定する．

[2]https://archive.org/details/stackexchange

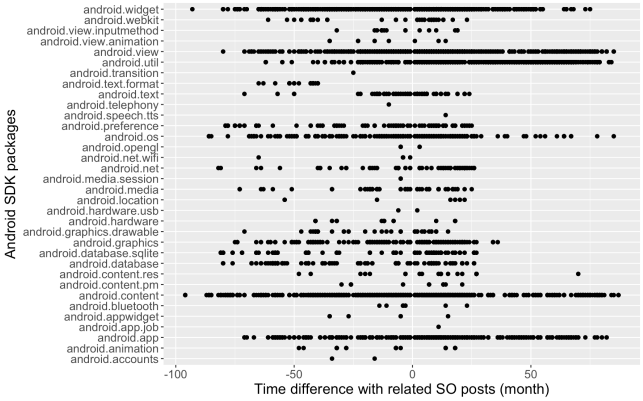

図 1: SDK のパッケージごとにみたソースコード変更時刻と SO 記事の投稿時刻との時間差

2.2.2 （RQ2）ソースコード差分は実際の SO 投稿に対して早く生成できるか

動機：SO において回答が投稿されるまでの時間的な課題を，ソースコード差分が改善できるかどうか探るため，SO 記事の投稿時刻に対するソースコード変更時刻の早さを調査する．

アプローチ：実際の SO 記事の投稿時刻に対するソースコードの変更時刻の差を計測する．ソースコードの変更時刻として，バージョン管理システムの Git でコミットが行われた時間を計測する．時間差の比較は，コミットが行われた時刻と SO 記事が投稿された時刻をそれぞれ Unix 時間に変換して比較する．

2.3 調査結果

2.3.1 RQ1

調査の結果，ソースコード差分と同じ Android SDK を含む SO 記事は 24,184 件であった．この数字は SO 投稿データセットの 47%を占める．また SO 記事と同じ Android SDK を使用するコード差分の件数は 7,692 件であった．この数字はコード差分のデータセット全体の 86%を占める．SDK のメソッド名を用いた粗い関連づけではあるが，データセットの半数近くの SO 記事に対して，コード差分は情報をもたらす可能性がある．

2.3.2 RQ2

ソースコード差分 8,929 件のうち，SO 記事と内容が近く，かつ SO 記事の投稿時刻より前に生成可能なソースコード差分は 2,203 件であった．この数字は，データセットのコード差分の 29%を占める．

本結果に対し，SO 記事の投稿時刻より前に生成可能なコード差分は，どういった SDK を使用したものなのかという疑問が起こる．そこで，SO 記事の投稿時刻に対するソースコードの変更時刻の差を，Android SDK のパッケージごとにみた結果を**図 1** に示す．図 1 中の一つの点は一対のコード差分を示している．

SDK のうち，特に android.database や android.preference，また android.text といった SDK は，SO 記事で議論される時刻より比較的早くコード差分で使用されており，android.util と android.app を使用したコード差分は，SO 記事で議論される時刻より比較的遅くコード差分で使用されていることが図から読み取れる．android.database や android.preference，また android.text の 3 パッケージに関する時刻の差は，中央値が-22，-7，-3 で，値の範囲は-80〜26，-79〜25，-71〜24 であった．一方で，android.util と android.app の 2 パッケージに関する時刻の差は，中央値が 45，8 で，値の範囲は-62〜84，-71〜82 であった．

前者のようなパッケージは，コード差分を生成する意義が大きいと考えられる．前者のようなパッケージはアプリケーションの基盤を，後者はアプリケーションの動作を扱うものであることがうかがえる．本結果は，基盤に関わる不具合対応はいち早く実施しなければならず，SO 記事がそれに追いついていない状況を示唆している．

3 まとめ

本稿では，SO における不十分性の改善を目的として開発者が API を使用したコード片に着目し，Android アプリケーションリポジトリから生成したソースコード差分が実際の SO 記事を充実する可能性を調査した．調査の結果，粗い粒度ではあるが，ソースコードのもつ可能性を関連付く SO 記事の量と時間の観点で見ることができた．今後の研究では，開発者が残したコメント文といったソースコード以外の情報についても，存在する SO 記事と関連付くか調査する予定である．

謝辞

本研究は，JP26240007，18H04097 による助成を受けた．

参考文献

[1] Chen, F. and Kim, S.: Crowd debugging, *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering*, ACM, pp. 320–332 (2015).

[2] F-Droid: Free and open source android app repository, [Online; accessed 2015-12-19] (2017).

[3] Mamykina, L., Manoim, B., Mittal, M., Hripcsak, G. and Hartmann, B.: Design lessons from the fastest q&a site in the west, *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, pp. 2857–2866 (2011).

[4] Nguyen, T. T., Pham, H. V., Vu, P. M. and Nguyen, T. T.: Learning API Usages from Bytecode: A Statistical Approach, *Proceedings of the 38th International Conference on Software Engineering*, New York, NY, USA, ACM, pp. 416–427 (2016).

[5] Parnin, C., Treude, C., Grammel, L. and Storey, M.-A.: Crowd documentation: Exploring the coverage and the dynamics of API discussions on Stack Overflow, Technical report, Georgia Institute of Technology (2012).

[6] Treude, C., Barzilay, O. and Storey, M.-A.: How Do Programmers Ask and Answer Questions on the Web?, *Proceedings of the 33rd International Conference on Software Engineering*, New York, NY, USA, ACM, pp. 804–807 (2011).

[7] Vasilescu, B., Serebrenik, A., Devanbu, P. and Filkov, V.: How social Q&A sites are changing knowledge sharing in open source software communities, *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, ACM, pp. 342–354 (2014).

[8] 西中隆志郎, 鵜林尚靖, 亀井靖高, 佐藤亮介:ソースコード変更履歴による Stack Overflow 記事の充実に向けて，情報処理学会 ソフトウェア工学研究報告 (2018 年 3 月).