ISS-P-57

クラッシュレポートの自動分類手法評価のための データセットの構築

師尾　彬† 　相澤　彰子†† 　浜本　隆之†
† 東京理科大学工学部電気工学科 　†† 国立情報学研究所

1. はじめに

クラッシュレポートとは，アプリケーションが強制終了した際のシステムの状況を記録したもので，多くの場合，スタックトレースの情報を含む．クラッシュレポートの自動分類では，報告されたクラッシュをバグごとにグループ化することで，開発者が頻度の高いバグを優先して修正することを可能にする．

自動分類の精度を高めることにより，より効率の良いバグ修正が可能となる．[1]において提案されている ReBucket 法では，前処理としてスタックトレースに存在するがクラッシュとは無関係な関数を削除する．ReBucket の適用では，事前に該当する関数のリストを準備する必要があるが，リストを作成するためには専門的な知識を持った開発者による頻繁な更新が必要であることから，クラッシュレポートと関数リストの両方を含むデータセットはこれまで構築されていなかった．そこで本稿では，クラッシュレポートの自動分類手法の評価を目的とする新たなデータセットの構築について報告する．

2. Mozilla におけるバグ修正の流れ

Mozilla では[2]によりクラッシュレポートが管理されている．報告されたクラッシュレポートは，スタックトレース上位のフレームから生成された Signature が付加され，これを元に分類される．開発者はクラッシュの傾向からバグレポートを作成し，バグ修正を行う．バグレポートと複数の Signature との関連付けを行うことにより，同一のバグに起因する複数の Signature を管理することができる．また，ある 1 つのバグについて複数のバグレポートが作成された場合に，一方の”DUPLICATE”(重複)として他方を関連付ける場合がある．また，Mozilla が現在運用している[2]においては該当する関数の正規表現のリストが公開されている[3]．

3. データセットの作成方法

構築するデータセットはクラッシュレポート群と，クラッシュレポートに紐づけられたバグの種類(すなわち分類の正解データ)の 2 種類のデータと関数のリストをもつ．

まず，最初に指定した範囲のバージョンで報告頻度の高い Signature の上位 300 件を取得する．各 Signature に関連付けられているバグレポートを取得し，さらにそれらの重複とされたバグレポートも取得する．重複とされたバグレポートに関連付けられた Signature と元のバグレポートの Signature は同一のバグに起因すると考えられるので，正解データではこれらの Signature は同一とみなすこととする．最後に各 Signature に属するクラッシュレポートをランダムに最大 100 件取得する．また，[6]と同様の形式に整形する．

4. データセットの概要

作成したデータセットの概要を表 1 に示す．クラスタ数は，正解データにおいて，クラッシュの原因と考えられるバグの数に相当する．ウェブブラウザのような大規模なアプリケーションでは，数多くのスレッドが実行されており，その多くがクラッシュと直接関係しない．そのため，取得したクラッシュレポート群のうち，クラッシュが発生したスレッドのスタックトレースのみをデータセットの対象とした．対象バージョンはデータセット作成時点の最新版である Firefox 48.0 のバージョンのうち，Nightly と Beta とした．作成したスクリプトとデータセットは近日中に公開予定である．

表 1. 作成したデータセットの概要

対象バージョン	48.0a1～48.0b99
クラッシュレポート数	37,146
バグレポート数	831
クラスタ数	727

5. おわりに

作成したスクリプトを用いると，バージョンを指定することで容易にデータセットを生成できる.今後は継続的にデータセットを作成する予定である．

参考文献

[1] Y. Dang et al., “ReBucket: A Method for Clustering Duplicate Crash Reports Based on Call Stack Similarity,” in Proceedings of the 34th International Conference on Software Engineering (ICSE), 2012, pp. 1084-1093

[2] https://github.com/mozilla/socorro

[3] https://github.com/mozilla/socorro/blob/master/socorro/siglists/irrelevant_signature_re.txt

[4] https://crash-stats.mozilla.com/home/product/Firefox

[5] https://bugzilla.mozilla.org/

[6] J. Campbell et al., “The Unreasonable Effectiveness of Traditional Information Retrieval in Crash Report Deduplication,” in Proceedings of the 13th International Conference on Mining Software Repositories (MSR), 2016, pp. 269-280.