6N-07

OSS開発プロジェクトにおける README ファイルの進化に関する理解

佐渡島 悠樹† 亀井 靖高† 佐藤 亮介† 鵜林 尚靖†

†九州大学

1 はじめに

GitHub は現在世界中の多くの開発者が使用しており，リポジトリによるバージョン管理とソースコードホスティングを行う開発プラットフォームである．OSS（オープンソースソフトウェア）プロジェクトの多くは，GitHub などのプラットフォームを介して開発を行い，成果物（ソフトウェア）をリリースする．

GitHub における README.md は開発者が新しいプロジェクトに関わる際，一番初めに表示されるドキュメントである．GitHub ではプロジェクトを作成する際、最初に README ファイルを作成し，どのようなプロジェクトなのか，プロジェクトの有用性，プロジェクトの使い方や誰がプロジェクトを作成し維持しているのかなどを記述することを推奨している．

しかしながら，開発者の中には README ファイルの作成に多くの労力を費やす人がいる．2017 年に行われた調査では，GitHub の利用者は README ファイルは重要であると考えているものの，内容が不完全であったり，古いままになっているものが多いため，読まれないことも多いという課題が指摘されている．

本研究では，REAMDE ファイルに関する上記課題を解決するために，ドキュメントの変更過程を調査することで README ファイルに記述すべき項目を明らかにし，ドキュメント作成の推奨方法の提案を目標とする．そうすることによって，Igor ら [1] の OSS の新規参入者に対する障害に関する研究におけるドキュメンテーション不足などが改善され，新規参入者のプロジェクトの立ち上げの手助けにもなると考えられる．

2 関連研究

Prana ら [2] は README ファイルの品質向上と関連情報の発見の効率化を図るために，ドキュメントの内容をセクションごとに自動分類し，ラベルづけするツールを設計した．彼らの作成したマルチラベル分類器は 0.746 の F1 値を達成し，有用性の評価に参加したソフトウェアの専門家の大部分は Prana らのツールが有用であると回答した．表 1 にカテゴリの分類内容を示す．

表 1: カテゴリ一覧 [2]

label	category	example section heading
1	what	introduction, project, background
2	why	advantages of the project, comparison with related work
3	how	getting started, how to run, installation, requirements, platforms, downloads, setup
4	when	project status, versions, project plans, roadmap
5	who	project team, community, mailing list, contact, acknowledgement, license,
6	references	API documentation, getting support, feedback, more information, translations, related projects
7	contribution	contributing guidelines
8	other	—

3 実験

本実験では，README ファイルの進化について理解するために，3 つの研究課題（以下 RQ：Research Question）を設定する．

RQ1 ドキュメントの初期バージョンには何が（どのカテゴリが）書かれているのか

RQ2 初期バージョンから最新バージョンまでの間に，どのカテゴリが追加されるのか

RQ3 変更回数とカテゴリの増減にどのような関係があるのか

3.1 データセット

プロジェクト. 本実験では Prana らの README ファイルの自動マルチラベル分類器の研究で使用されたデータセットを利用する．これは GitHub 内のプロジェクトをランダムにクローンし，その中から README ファイルにタイトルしか書かれていないと思われるもの等の実験に不

Understanding the evolution of README files in OSS projects

Yuki Sadoshima†Yasutaka Kamei†Ryosuke Sato†Naoyasu Ubayashi†

†Kyushu University

sadoshima@posl.ait.kyushu-u.ac.jp

{kamei, sato, ubayashi}@ait.kyushu-u.ac.jp

表 2: 初期と最新バージョンでのカテゴリ

label	初期時の件数	追加件数	最新時の件数
1	122	239	361
2	26	71	97
3	111	238	349
4	19	63	82
5	50	160	210
6	49	177	226
7	25	78	103
8	6	20	26

適切なプロジェクトを除外した 404 のプロジェクトが含まれている．

404 件のうち，README ファイルが英語以外の言語で書かれている等でカテゴリ分類されていないものがある．本研究では，分類されていないドキュメントを除いた 370 件について調査した．

データセットの取得．先行研究である Prana らのデータセットには，最終バージョンのカテゴリは記載されているものの，初期バージョンについては含まれていない．Prana らは，README.md を＃がついている部分（見出し）ごとにセクションとして分け，セクションの中にどのような内容が書かれているのかラベル付けした csv ファイルを作成している．csv ファイルの中にはプロジェクト名，見出し，ラベルが記述されている．記述されているプロジェクト名を利用して GitHub からプロジェクトをクローンした．

初期バージョンのカテゴリ分類．READEM ファイルの変更情報を得るために，“git log” コマンドを利用した．これによって初期バージョンのコミット ID を取得した．取得したコミット ID を利用して README ファイルを初期バージョンに戻し，初期バージョン時にある見出しを抽出した．抽出した見出しと csv ファイルの見出しを見比べ同じものがあれば初期バージョンの README ファイルにも同じカテゴリの内容が含まれていると考え，それを元に初期バージョンのファイルにもカテゴリのラベル付けを行った．

変更回数と追加カテゴリ．初期バージョンのコミット ID を取得する時と同様，”git log”コマンドを利用してコミット回数を調査した．また，初期バージョンのラベルと csv ファイルのラベルを見比べることによって，追加されたカテゴリ数を調査した．

4 実験結果

RQ1．表 2 は初期バージョンでのカテゴリ数と最新バージョンまでに追加されたカテゴリ数と最新バージョンでのカテゴリ数を記述している．なお，初期バージョンではどのカテゴリも含まれていないドキュメントが一番多く，211 件であった．

表 3: コミット数と追加カテゴリ数の関係

追加カテゴリ数	件数	コミット数の平均
1	132	15.0
2	51	16.7
3	77	20.5
4	59	38.1
5	47	63.3
6	28	123.0
7	9	59.7
8	1	80.0

表 2 に示すように，最新バージョンでも多い what（label:1）と how（label:3）が，それぞれ 122 件と 111 件であり，初期バージョンでも多いことがわかる．

RQ2．表 2 に示すように，who(label:5) と reference(label:6) がそれぞれ 160 件と 177 件であり，途中で追加される割合が高い．追加件数としては what と how の方が多いが，2 つのカテゴリは初期バージョンから多く，途中で追加される割合としてはそれぞれ 66 %と 68 %であり，who と reference の方が追加される割合はそれぞれ 76 %と 78 %で高い．

RQ3．表 3 は，初期バージョンから最新バージョンまでで追加されたカテゴリ数別にその件数とコミット数の平均を記述した．コミット数の平均に着目すると，コミット回数が多い方が追加カテゴリ数も多くドキュメントの内容が充実していると考えられる．追加カテゴリ数が 7,8 個の場合は件数が少なく，偏った値になっている．

5 おわりに

本稿では README ファイルの初期バージョンからの進化に着目して調査を行った．今後は複数カテゴリを持つドキュメントはどのような順番で追加されるのかや，ソースコードやマニュアル内で頻出する単語と README ファイルの関連性などに着目して調査していき，ドキュメントの記述方法の提案を目指す．

謝辞．本研究は，中島記念国際交流財団による助成を受けた．

参考文献

[1] Steinmacher, I., Conte, T. U., Treude, C. and Gerosa, M. A.: Overcoming open source project entry barriers with a portal for newcomers, *Proceedings of the 38th International Conference on Software Engineering, ICSE 2016, Austin, TX, USA, May 14-22, 2016*, pp. 273–284 (2016).

[2] Prana, G. A. A., Treude, C., Thung, F., Atapattu, T. and Lo, D.: Categorizing the Content of GitHub README Files, *arXiv preprint arXiv:1802.06997* (2018).