データマイニング手法を用いたOSS開発プロジェクトの分類調査

プロジェクトマネジメントコース　矢吹研究室　1142009　安藤　勇樹

1.　研究背景

　ホスティングサービスであるGitHubでは様々なソフトウェアが開発されている．2013年12月にはGitHub.com上に1000万件のリポジトリが作成され，ユーザ数は400万人を超えた．数多くのプロジェクトが公開されているGitHubを調査すれば，ソフトウェア開発プロジェクトの分類が可能であると考えられる．

　過去にGitHub上のプロジェクトのチケットを調査し，プロジェクトを分類するという試みがあった．[3]この研究では，分類の解釈を人間が主観的に行っており，客観性に欠けていた．そこで，本研究ではデータマイニング手法を用いて分類を客観的に行う．

　GitHubの特徴として，Issueを利用してのチケット駆動開発の運用が挙げられる．チケット駆動開発とは，チケット管理から生まれたプロジェクト管理技法の一つである．ソフトウェア開発中に発生した課題や変更の内容をチケットという形で発行してから開発するという手法である．どのような課題が発生し，どのような変更を行ったのかを可視化できるため，進捗管理が容易となる．

　チケット駆動開発の運用を中心としたプロジェクトは，未消化チケット数と消化済チケット数の時系列変化から進捗状況を判断できる．未消化チケット数の増加は何らかの課題が発生した事を示し，消化済チケット数の増加は課題の解決を示している．つまり，この2種類のチケット数の差が小さければ進捗状況は順調であり，差が大きければ進捗状況は不順であると判断できる．

2.　研究目的

　GitHub上のプロジェクトを対象とする．チケット数の時系列変化に着目し，データマイニング手法を用いてプロジェクトの分類を定量的に行う．

3.　研究方法

　チケット（GitHub上ではIssue）をGitHub内のスター数ランキング上位50件のプロジェクトからAPIを用いて収集する．収集するチケットは未消化チケット（GitHub上ではOpenIssue）と消化済チケット（GitHub上ではCloseIssue）の2種類である．この2種類のチケットの時系列変化を調査し，プロジェクトの分類を行う．分類には統計解析ソフトRを用いて定量的に行う．

4.　成果物のイメージ

　50件のプロジェクトのIssueデータをRで時系列解析し，4パターンに分類する．

5.　進捗状況

　50件のプロジェクトのIssueデータをRで時系列解析した．利用した解析手法は階層クラスター分析と非階層クラスター分析である．変数には標準化した多項式の次式を指定した．これはIssueデータを折れ線グラフ化した際に追加できる近似曲線である．以下の図は解析した結果である．

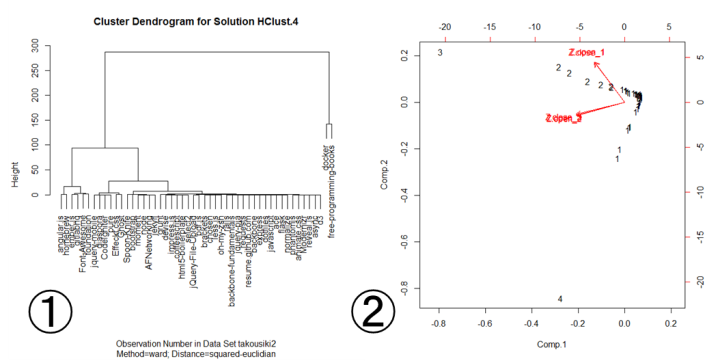


図 1. OSS開発プロジェクトの分類化

　①は階層クラスター分析，②は非階層クラスター分析の結果である．Issue数の増加率が高いパターンが6件．増加率が低いパターンが42件．増加率が非常に高いパターンが1件．CloseIssueの発生が非常に速いパターンが1件．計4パターンに分類することができた．

6．今後の計画

　プロジェクトを4つのパターンに分類できたが，どのような定義で分類されているのかは完全に把握できていない．今後は分類の定義を明確にする．

　参考文献

[1] 小川明彦, 酒井誠. チケット駆動開発. 翔泳社, 2012-8-23

[2] 大塚弘記. GitHub実践入門~Pull Requestによる開発の変革~. 技術評論社, 2014-4-25

[3] 久保孝樹. “調査結果考察”. 「チケットを活用するオープンソースソフトウェア開発の実態調査」, 千葉工業大学（未公刊）, 2013, p. 42-66.

ファイルは、「/卒業論文/2013/安藤勇樹」に、ファイル名は「卒研中間審査用研究概要.docx」とする。

①GitHubとは何か、②そこで何を目指すのか、③そのための試みとして過去にどういうことがなされてきたか、④それを踏まえて何を考えているか、⑤現状はどうか、⑥今後はどうするか  
君は③から書き始めている。

久保の卒論は参考文献に入れる。文献リストにおける卒論の書き方は、[https://web.archive.org/web/20130215142017/http://sti.jst.go.jp/sist///handbook/sist02\_2007/sist02.htm#5-4](https://web.archive.org/web/20130215142017/http:/sti.jst.go.jp/sist/handbook/sist02_2007/sist02.htm#5-4) を参照。

「チケット」が定義されていない。

「チケット駆動開発を運用して成果物を作成しているプロジェクトを対象とする」とあるが、チケット駆動開発をしているかどうかは、調べてみなければわからない。

「R」はRでよい。

「¬ ラベル1はcomp.1とcomp.2・・・」の意味がわからない。スペースが足りなければ、記述のレベルを変えるように。