ビッグデータ解析ソフトウェアを用いたGitHubデータマイニング

プロジェクトマネジメントコース　矢吹研究室　1142046　小池 由也

1．研究背景

ソフトウェア開発プロジェクトのための共有ウェブサービスである，GitHubのプロジェクトについて調べれば，オープンソフトウェア開発プロジェクトの実態がつかめるはずである．

実際にGitHubを調べて分かったことの例として，怒りの表現を含むコミットメッセージの割合，地域によるオープンソースプロジェクトへの貢献者などの分布図などがあげられる．これらの結果は，GitHub Data challengeというイベントで上位に入賞している分析結果である．

GitHubのデータ解析は難しい．なぜなら，データが膨大なため，その収集と処理が難しいからである．データの収集が難しいという問題は，一つのプロジェクトにより簡単になった．大量のデータを集めるために，GitHubのプロジェクトのタイムラインを記録し，アーカイブ化させ，簡単にアクセスできるようにするためのプロジェクトでGitHub Archiveである．

データの処理が難しいという問題は，データ量が多すぎるために膨大な量のデータを処理するソフトウェアが少ない点である．GitHub Archiveと連動させデータ処理ができるソフトウェアにGoogle BigQuearyがある．BigQuearyは，簡単にビッグデータを処理するためのソフトウェアであり，SQLに似たクエリを従来のやり方よりも短時間で簡単に実行できる．このソフトウェアの登場により手軽に大量のデータを処理することができるようになった．

これまでの調査で，オープンソフトウェア開発でどのようなプログラミング言語がよく使われているかを調べることに成功したが， プロジェクトがForkされる確率の，プログラミング言語による違いが分かれば，オープンソフトウェア開発プロジェクトについての理解が深まると思われる．Forkとは，GitHub上で公開されている成果物に独自の変更を加える際に行う複製のことである．

2．研究目的

GitHub上で公開されているオープンソフトウェア開発プロジェクトをGoogle BigQuearyを利用し調査する．オープンソースソフトウェアの開発プロジェクトにおいて，使用するプログラミング言語が異なると，Forkされる確率，つまりプロジェクトに貢献する人が現れる確率が異なるということが課題研究の段階で分かった．

しかし，この結果は，Forkされた回数が多いものについてのみ調査して得られたものであった．

そこで本研究では，Forkされた回数が非常に少ないものも対象にして，プログラミング言語による貢献者の出現確率を調査する．

3．研究方法

大量のデータを処理することが予想されるのでGoogle BigQuearyを利用する．Google BigQuearyを使って，GitHub上のプロジェクトが採用しているプログラミング言語とForkされている数を集める．BigQuearyで集めたデータをMySQLを用いて平均，分散，ヒストグラムといった形で視覚的に違いが分かるよう処理を加える．

4．結果

Google BigQuearyを利用して2012年2月12日から2014年11月18日までのGitHubのタイムラインデータを集めた．集めた約20万件プロジェクトのデータをMySQLで整理した結果の一部が以下の表である．



表1 言語ごとの平均Fork数と標準偏差

5．考察

オープンソースソフトウェア開発の現状はやはりJavaScriptとRubyを用いた開発が多く行われているようである．また，プロジェクト数が少ない言語ではFork数も減りForkされる確率も減少するということが分かった．

参考文献

[1] Brian Doll. The GitHub Data Challenge

<https://github.com/blog/1118-the-github-data-challenge> (参照2015-1-26).