# 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わかっている．

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 されている数を収集・統計処理し，Fork される確率のプログラミング言語による違いを明らかにする．

## 4. 成果物のイメージ

GitHub のプロジェクトで使われているプログラミング言語を解析し，言語による Forked 数を統計処理する．それにより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予想される．

## 5. 進捗状況

GitHub Data Challenge の入賞者のデータ解析の手法を調べ，プロジェクトの貢献者の分布図や GitHub での活動を可視化させプロジェクトの動きなどがわかった．

これまでの調査の際に実際に Google BigQueary を使用し GitHub に登録されているプロジェクトで使われているプログラミン言語を調べることができた．

## 6. 今後の計画

Google BigQueary を活用し，そこで出た結果を様々な手法を利用し，プログラミング言語による貢献者の出現確立を調査する．

## 参考文献

[1] The GitHub Data Challenge　2012-5-1.
https://github.com/blog/1118-the-github-data-challenge