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にあげられている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，誰でも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誰でも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よく使われているプログラミング言語を調べることに成功したが， よく使われているプログラミング言語のForked数の違いが分かれば，オープンソフトウェア開発プロジェクトの実態を明らかにできると思われる．Forkとは，GitHub上で公開されている研究成果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の実態をGoogle BigQuearyを利用し調査する．GitHub Data Challengeで全言語共通で人気のForkはあるのだが，各言語では人気の Forkは上位5つまでのリスト化しか行われていないようなので，これまでの調査でわかったTop LanguageのFork数を可視化する．それにより，数値だけで調査していたソフトウェア開発プロジェクトの状態を色々な視点から実態調査する．

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で集計をとった結果を統計処理し，プロジェクトの実態を調査し数値以外で分かるプロジェクトの実態を見つける．

4．成果物のイメージ

　GitHubにあげられているプロジェクトの内，最も使われている上位20件の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．それによってオープンソフトウェア開発の実態が視覚的にわかると予想される．

5．進捗状況

　GitHub Data Challengeにあげられている入賞者のデータ解析の手法を調べ，実際にどういうことがわかるのか，地域ごとのプロジェクトの貢献度やプログラミング言語によるリポジトリの活動内容などがわかった．

これまでの調査の際に実際にGoogle BigQuearyを使用しGitHubで使われている人気プログラミング言語を調べることができた．

各プログラミング言語のForked数を調べるためのコードの作成が完了したので，これからそれを使い実際にForked数を調査する．

6．今後の計画

　Google BigQuearyにコードを入れ，そこで出た結果を様々な手法を利用し可視化させ数字だけでは分からなかった結果を見られるようにする．

参考文献

[1] The GitHub Data Challenge　2012-5-1.

<https://github.com/blog/1118-the-github-data-challenge>

[2]The GitHub Data Challenge Ⅱ 2013-4-3.

<https://github.com/blog/1450-the-github-data-challenge-ii>