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，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というものがある．

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というものがある．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使われている人気のプログラミング言語は調べることに成功したが，人気だけしかわかっていないのでプログラミング言語による違いやPopular Forked数などが分かれば，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．

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Big Data処理技術を利用し調査する．GitHub Data Challengeで全言語共通でPopular Forkedはあるのだが，各言語ではPopular Forked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Forked数を可視化することにより，他の視点からソフトウェア開発のプロジェクトを見ることで違った視点からのソフトウェア開発の実態を調査する．

3．研究方法

これまで調査したTop LanguageのForked数を調べるにあたって，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表計算ソフトでまとめ．プロジェクトの実態を調査する．

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数を可視化する．それによってオープンソフトウェア開発の実態が明確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の手法を調べ実際にどういうことがわかるのかを調べた．

これまでの調査の際に実際にこのソフトウェアを使用し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収集する．

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りようし可視化させ異なる視点から調査結果を見れるようにする．

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>

[3]Data Challenge Ⅱ Results　2013-6-26.

<https://github.com/blog/1544-data-challenge-ii-results>