オープンソフトウェア開発プロジェクト解析

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

パソコンやスマートフォンの普及により普段の生活で扱うデータも大きくなった．そのため，より早く簡単にデータを処理するソフトが生まれた．

Googleでは，誰でも簡単にビッグデータを処理するためのソフトGoogleBigQueryというものがある．GoogleBigQueryとは，Googleの提供しているサービスであり，SQLに似たクエリを数秒で簡単に実行できる．このソフト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．

大量のデータを集めるものとして，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．これは，2012年から始まったプロジェクトでGitHubチームとBigQuearyチームが協力して進めているサービスである．

GitHub Archiveの登場により誰でも簡単にGitHubのデータを分析できるようになった．それによりGitHubでは，データ統計のイベントでGitHub Data Challengeというものが2012年から始まった．

このGitHub Data Challengeは，GitHubで公開されているプロジェクトを集めて解析し結果を可視化しそこから開催者が良いと思ったものに投票する企画である．上位入賞者の分析結果を見て分かることは，プロジェクトにおいて進捗状況やプロジェクトに参加している人の心境の確認ができる．

分析結果の例として，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．

このように，GitHubで集めたデータを基に解析を行え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だ．

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課題研究で調べたTop Languageの各言語までのForked数を可視化することによりソフトウェア開発のプロジェクトの実態を調査する．

3．研究方法

課題研究で調査した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可視化を調べ実際にどういうことがわかるのかを調べた．

GitHubで最も使われている上位20件のプログラミング言語の収集をする．そこから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

　まだ、上位20のプログラミング言語の収集が終わった段階なので，これからGoogle BigQuearyにコードを入れ，そこで出た結果を可視化する．

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>