**ビッグデータ処理技術を活用したオープンソフトウェア開発プロジェクト解析**

PMコース　矢吹研究室　1142046　小池由也

1. 研究の背景

「ビッグデータ」という言葉が流行することからもわかるように，私たちは膨大な情報を生み出しながら生活している．

　ビッグデータは「量（Volume）」「多様性（Variety）「速度（Velocity）」の「3V」で特徴づけられる[1]． GoogleやYahoo!に寄せられる大量の検索クエリや， Twitter，FacebookなどのSNSに投稿される文章や画像，動画，スマートフォンを利用するサービス等で収集される位置情報データ，防犯カメラで記録される人間の表情や動きのデータすべて，「3V」の特徴を持っている．

　さまざまな場所で，これらのビッグデータをビジネスに活用することが試みられている．企業でもビッグデータを活かして，商品の購買履歴やサイト内のアクセス情報を基に，商品を購入する際に他のおすすめの商品やバーナー広告の表示を行っている[2]．他にも利用者の店内の行動データを基に利用者の目に留まりやすい所に商品を陳列させて売り上げを20％増加させた企業もある[1]．

　ビッグデータをこのように活用することを可能にしたのは，インターネットの普及とコンピュータの性能向上，解析アルゴリズムの発展である．現在では，インターネットの普及とコンピュータの処理技術の向上により「3V」のデータと相性の良い分析法・アルゴリズムが発達し必要な情報の蓄積と解析が以前に比べて簡単に行えるようになりデータの使い道が「法則を確かめる道具」から「法則を発見する道具」として使われるようになった[3]．そのため従来とは異なり精度ではなく量を重視し因果関係ではなく相関関係が重視されている．

　ソフトウェア開発の現場でも，ビッグデータが生み出されている．プロジェクトでの色んなドキュメントや全ての段階の草案，アップデート履歴などのたくさんのデータも集まればビッグデータでそれが集まっているバージョン管理システムの一種がGitHubである．

1. 研究の目的

オープンソースソフトウェア開発で生み出されるビッグデータを解析し，オープンソースソフトウェア開発の実態を明らかにすることを目指す．サンプリングにより調査されていたオープンソフトウェアの実態を，ビッグデータ処理技術を利用し調査する．矢吹研究室ではGitHubの研究は以前にも行われているが，主にサンプリングによる調査である．ビッグデータの処理技術を利用し調査することで，サンプリング調査ではわからなかった実態が明らかになることが期待できる．

1. プロジェクトマネジメントとの関連

オープンソフトウェア開発プロジェクトがどのようにマネジメントされているかを調査する研究である．

1. 研究の方法

　本研究では，GitHub ArchiveとGoogle BigQuearyを主に利用する．

　GitHub Archiveは，GitHubのタイムラインを記録しアーカイブ化し簡単にアクセスできるようにするもので，これを利用しGitHubに上げられてる今回の研究で必要なデータを集める．

　Google BigQueryは，大量のデータに対して高速にクエリを実行可能なGoogleのサービスで，クラウドにあるビッグデータをSQLを使って分析できる．GitHub Archiveで集めたGitHubのデータの分析管理を行う．

　これらの技術を用いて，以下のように研究を進める．

* 1. ビッグデータがどのようなものか，またどのように利用されているのかを調べる．
  2. GitHubがどのようなもので，どのようなプロジェクトが存在しているのかを調査する．
  3. GitHubの大量データを利用できるようにする（GitHub Archiveも活用する）．
  4. 処理技術のテストとしてプログラミング言語の統計を取る．
  5. 矢吹研究室でサンプリングにより調査されているもの（ソフトウェアの規模・ユーザの行動統計・Issuesの利用統計）の調査規模を拡大する．

1. 現在の進捗状況

図1　プログラミング言語レポジトリ数

例として，GitHubで2011/1/1～2012/11/15作成されたレポジトリを対象に，プログラミング言語ごとのレポジトリを調査した．図1は，この期間内に使われているプログラミング言語の頻度である．このグラフから，GitHubで公開されているオープンソフトウェアの開発に使われているプログラミング言語は，JavaScriptが圧倒的に多いことがわかる．

1. 今後の計画

|  |  |
| --- | --- |
| 計画表 | |
| 内容 | 日程 |
| データの収集・解析ツールの試運転 | 2014年1月～ |
| データの収集 | 2014年4月～ |
| データの解析 | 2014年6月～ |
| 解析結果の有用性の調査 | 2014年8月～ |
| 卒業論文執筆 | 2014年10月～ |

参考文献

[1] 城田真琴．ビッグデータの衝撃－巨大なデータが戦略を決める，ｐ．22-25．東洋経済新報社．2012

[2] WORKSIGHT“誤解されている「ビッグデータ」の本質”．2013-3-15．

<http://www.worksight.jp/issues/287.html>

[3] ニッセイ基礎研究所“ビッグデータで何が変わるか？”．2013-11-8．

<http://www.nli-research.co.jp/report/report/2013/11/repo1311-c3.html>

[4] Google Developer RelationsJapanBlog“Google BigQueryを使ってGitHubのデータを研究しよう”．2012-6-1．

<http://googledevjp.blogspot.jp/2012/06/google-bigquery-github.html>