**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

　ここ数年で，情報を得るためにウェブサイトを検索する機会が増えた．最近では「ビッグデータ」という言葉も聞くようになり，改めて大量の情報に囲まれていることが確認できる．

　ビッグデータという言葉には明確な定義がないが，私の研究では「量（Volume)」「多様性（Variety）「速度（Velocity)」の「3V」で特徴づける[1]．インターネットの普及によりGoogleやYahoo!の検索欄には，この瞬間にも多くの人がさまざまな言葉を入力している．さらにTwitter，FacebookなどのSNSには画像やつぶやきが投稿されていて高速で多様かつ大量のデータが生み出されるようになった．また，スマートフォンから取得される位置情報データや，防犯カメラで記録される人間の表情や動きのデータも「高速で多様かつ大量に生み出される」の「3V」の特徴に当てはまる．

　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]．

　現在では，インターネットの普及とコンピューター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．

　ビッグデータの処理技術を利用することで，以前まで量が多すぎてサンプリング調査で済まされていたものがサンプリングする必要なく，すべてのデータを利用して調査できるようになった．

　矢吹研究室ではオープンソースのソフトウェアリポジトリであるGitHubでサンプリング調査がいくつか行われているので，GitHubに上げられているソフトウェア開発を調査対象とする．

1. 研究の目的

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なので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
   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利点，欠点とビッグデータが利用された事例を調査する．
   2. オープンソースプロジェクトを行う場を提供しているGitHubがどのようなものなのか，どのようなプロジェクトがどれほど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. 例としてプログラミング言語の統計を取ってみる．
   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の規模を拡大する．
2. 現在の進捗状況

図1　プログラミング言語レポジトリ数

今回の私の研究では，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．

GoogleBigQuery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のでGitHub Archiveで集めたGitHubのデータを分析管理を行う．

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，期間内に使われているプログラミング言語のグラフである．

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>