Wikipedia の履歴データ解析

PM コース　矢吹研究室　1242005　石井康之

1　研究の背景

Wikipedia はウェブ上で公開されている誰でも自由に編集できる有名な百科事典のサイトである [1]．ウェブでの検索結果では 1 ページ目にあげられるほど使用頻度の高いサイトで，現在では約 10 年も続いているほどの古参サイトだ．Wikipedia では長年の編集履歴データを公開して，誰でも取得できるようにしている．しかし，そのデータの量は個人で取得し扱うにはあまりにも膨大である．このデータをいわゆるビッグデータというものではないかと考えた．

ビッグデータとは市販されているデータベース管理ツールや従来のデータ処理アプリケーションで処理することが困難なほど巨大で複雑な データ集合の集積物を表しているものである [2]．適切なハードウェアと環境を用意しないと，時間と費用が多くかかるものだ．さまざまな企業がビッグデータの活用の必要性を求めている．YAHOO では，2014 年 12 月 14 日に行われた予定の衆議院選の議席予測を行った [3]．他にも，ビッグデータをマーケティングに活用することを考えられている [4]．

Wikipedia を 1 つのプロジェクトとみなして，どのように成功していったのかを編集履歴データを解析し調査するために，ビッグデータというものを解析できる技術を取得することを考えた．

2　研究の目的

ビッグデータを解析できる技術を取得することを目指す．Wikipedia の膨大な編集履歴データを扱えるようにするためである．

3　プロジェクトマネジメントとの関連

Wikipedia を 1 つのプロジェクトとみなすと，品質管理マネジメントと人的資源マネジメントが関連付けられる．

品質管理マネジメントに関連付くと考えられるのは，オープンな共同作業のプロジェクトにおいて，悪意のある編集の荒らしや，研究の背景で言った編集合戦があるにもかかわらず，現在では誰もが信頼して使うような百科事典になったためである．

人的資源マネジメントに関連付くと考えられるのは，多くのボランティアの人々の協力により，Wikipedia が多くの情報を持つ百科事典になったためである．

4　研究の方法

当研究では XML dump[5] と Google BigQuery[6] を利用する．

XML dump とは，Wikipedia のすべての記事の完全な編集履歴を提供しているものである．これを利用して研究で必要な Wikipedia であげられているデータを取得する．

Google BigQuery とは，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 を使って解析できる．XML dump に入っている編集履歴からデータ解析を行う．

これらを用いて，以下のように研究を進める．

① BigQuery の利用の仕方を調べる [7]．

② BigQuery を用いて Wikipedia の作業履歴データを取得する．

③ BigQuery を用いて Wikipedia の作業履歴データを利用できるようにする．（XML dump からデータを抽出し，解析できるようにもする）

④ BigQuery の処理技術のテストとして記事ごとの差し戻し履歴の統計を取る．

⑤ BigQuery を用いて抽出したデータと，差し戻し履歴の統計を参考にしているサイトを照らし合わせて一致して

いるか確認する．

5　現在の進捗状況

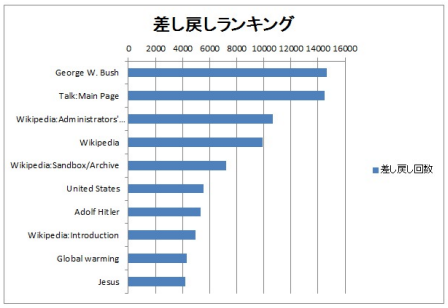
以下のように進んでいる．

① BigQuery を用いて Wikipedia の履歴データを解析した．

② BigQuery を用いて履歴データの中から差し戻し履歴の部分を記事ごとに抽出し，ランキングを作成した．

③差し戻し履歴の統計が参考にしているサイトと照らし合わせたら一致しなかったので，原因を調査した．

④扱っていたデータの範囲はほぼ同じだった．差し戻し履歴の抽出方法が違っていたと考察する．



# Rv	Article
1447	Feces
1404	Vagina
1368	Edit
1322	Fat
1311	Sex
1306	Idiot
1298	Intellectual disability
1282	America
1279	Midget
1268	Penis

図 1　差し戻し比較

6　今後の計画

以下のように進める予定である．

① Wikipedia の全編集履歴データをダウンロードする．

②引き続き BigQuery，または API を用いて解析する．

参考文献

[1] アンドリュー・リー. ウィキペディア・レボリューション. ハヤカワ新書, 初版, 2009.

[2] ビッグデータ - wikipedia.

[3] Yahoo!ビッグデータ議席予測　自民は 300 超、民主は 60 台.

[4] ビッグデータ活用マーケティング、最初の一歩とは?

[5] ウィキメディア財団による全プロジェクトのデータベース・ダンプ.

[6] Google cloud platform bigquery.

[7] What is bigquery? - google cloud platform.