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 はウェブで有名な 1 つのフリー百科事典である．ウェブで何かを検索したときに 1 ページ目にあげられるほどである．このフリー百科事典は，その名のとおり誰でも自由に編集できる百科事典というので有名になった．自由に編集できるというのは，善意を持って編集をする人以外にも，悪意を持って編集をする人も稀にいる．それはこの Wikipedia も例外ではない．内容の信頼性を疑問視する声もある [1]．しかし，私たちが使用しているときにそのような影はめったに見ない．現在の Wikipedia は約 10 年も続いているほどの古参サイトである．

Wikipedia ではその長年の編集履歴データを公開し，誰でも取得できるようにしている．そのデータ量は個人で取得し扱うにはあまりにも膨大で，頭を悩ませるものである．それらはいわゆるビッグデータという．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].

最近では，さまざまな企業がビッグデータの活用の必要性を叫んできている．YAHOO では，近頃行われる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].

2　研究の目的

Wikipedia の履歴データを扱えるようになるために，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 の一部の実態を，ビッグデータ処理技術を利用し解析する．今回は Wikipedia の編集で差し戻しをされている部分のデータについて取得する．参考にしているサイトと同じデータを取得できるようにする [5].

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フリー百科事典になったため．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フリー百科事典になったためである．

4　研究の方法

本研究では，XML dump と Google BigQuery を利用する．

XML dump 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 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．

これらの技術を用いて，以下のように研究を進める．

・BigQuery の利用の仕方を調べる [6].

・Wikipedia の作業履歴データを取得する．

・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）

・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．

・差し戻し履歴の統計を参考にしているサイトで出していたので，照らし合わせて一致したのが取得できたか確認

する．

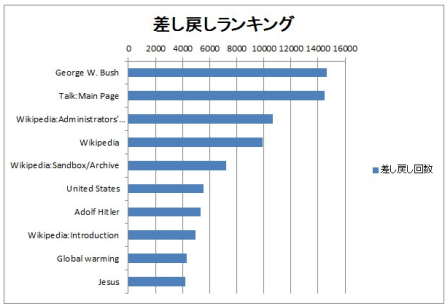
5　現在の進捗状況

・Wikipedia の履歴データを，BigQuery を用いて解析した [7].

・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.

・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.

・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

・Wikipedia の全編集履歴データをダウンロードする [8].

・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] Wikipedia statistics - edit and revert trends: English.

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

[7] Google cloud platform bigquery.

[8] 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.