ビッグデータ処理技術を用いた Wikipedia マイニング

PM コース　矢吹研究室　 1242005　 石井康之

1　研究の背景

Wikipedia はオープンなプロジェクトにおける最も有名な成功事例のひとつである [1]．使用頻度が高いサイトで，約 10 年も続いているほどの古参サイトである．また多言語で展開されており，2014 年 12 月 1 日で 288 言語も開設され，記事の数なら総合計で 3000 万件を超えた [2]．Wikipedia は誰でも自由に編集できる百科事典なので，内容の信頼性を疑問視する声もある．問題のある記述がなされた場合，それは善意のある人に一任される．完全な自由主義なため悪意のある人の編集を防ぎきれないという指摘がある．記事は完成・確定されることはないので，新しい情報にいつでも改変することができる．

Wikipedia について調査することで，オープンなプロジェクトのマネジメントについての知見が得られることが期待される．誰でも自由に編集できる状況において，どのように品質管理がなされているのか調査する．また多くの執筆者の協力によって成り立っているプロジェクトなので，どのように人的資源が活用されているのか調査する．

Wikipedia のデータは膨大なため，調査のためにはビッグデータを扱う技術が必要である．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だ [3]．適切なハードウェアと適切な環境を用意しないと，個人で扱うには時間と費用が多くかかってしまう．

そこで本研究では，Wikipedia を調査できるようなビッグデータ処理技術を調査し，そのベンチマークを行う．この調査により，オープンなプロジェクトにおけるマネジメントのあり方について調査するための技術を身につける．

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の 2 つを関連付けられる [4].

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，ふさわしくない投稿を繰り返し続ける行為の荒らしや，話し合いをせず他者の編集を繰り返し差し戻しを行う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

本研究では 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．これを利用して 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 [8] 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.

表 1　差し戻し回数の比較　両記事とも回数が多い順に出している

本研究での抽出結果

ランキング	記事	差し戻し回数
1	George W. Bush	14653
2	Talk:Main Page	14498
3	Wikipedia:Administrators' noticeboard/Incidents	10643
4	Wikipedia	9906
5	Wikipedia:Sandbox/Archive	7234
6	United States	5540
7	Adolf Hitler	5328
8	Wikipedia:Introduction	4976
9	Global warming	4314
10	Jesus	4196

参考文献

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

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.
3. Wikipedia の各記事の差し戻し回数の集合知を描き，どのような傾向があるか調査する.

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, 2009.

[2] ウィキペディアン.Wikipedia:全言語版の統計.Wikipedia. http://urx2.nu/fhg0(参照 2014-11-4).

[3] ウィキペディアン. ビッグデータ.Wikipedia. http://urx2.nu/fhey(参照 2014-10-15).

[4] Project Management Institute. プロジェクトマネジメント知識体系ガイド (PMBOK ガイド). Project Management Institute, 第 5 版, 2013.

[5] ウィキペディアン. 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. WikipediaDownloads. http://dumps.wikimedia.org/(参照 2014-10-14).

[6] Google.BigQuery.Google Cloud Platform. https://cloud.google.com/bigquery/(参照 2014-10-14).

[7] Google.What is BigQuery?.Gogle Cloud Platform. https://cloud.google.com/bigquery/what-is-bigquery?hl=ja(参照 2014-10-7).

[8] Erik Zachte.Edit and Revert Trends: English.Wikipedia Statistics. http://stats.wikimedia.org/en/editsrevertsen.htm(参照 2014-10-9).