ビッグデータ処理技術を用いた Wikipedia マイニング

プロジェクトマネジメントコース・ソフトウェア開発管理グループ　矢吹研究室　1242005　石井康之

1. 序論

当研究では，Wikipedia の全編集履歴をデータマイニングすることによって，Wikipedia の管理者の動向を調査する．

ビッグデータとは，市販されているツールや従来のデータ処理で行うことが困難なほど巨大なデータ集合の集積物のことである．

Wikipedia とは，非営利の Wikimedia 財団がインターネット上で運営する，無料のオンライン百科事典プロジェクトである．このプロジェクトは匿名のボランティアの人々の協力によって日々動いている．Wikipedia の記事の品質が保たれて，現在も動いている理由として，管理者の存在が影響しているのではないかと考えた．

Wikipedia では，プロジェクトが開始してから今までの全編集履歴データを提供している．このデータを扱うためにビッグデータ処理技術を用いることが出来る．

そこで当研究は，多言語版 Wikipedia もデータ解析できる技術を得るために，ローカルで研究が行える環境を用意し，データマイニングを行う．

2. 目的

Wikipedia を一つのプロジェクトとみなし，このオンライン百科事典で管理者の動向がどのように変化しているか調査する．

また，Wikipedia の編集履歴データを扱う際に必要なツールやプログラミングソースについての知識と，ビッグデータを扱う処理技術を得る．

3. 手法

以下のとおり研究方法を行う．

1. Wikipedia 日本語版の編集履歴まで含んだファイルをダウンロードし，ローカルでデータマイニングを行う．
2. Wikipedia の管理者の動向を解析する．
3. 管理者の動向がプロジェクトの動きにつながっているのか調査する．

4. 結果

日本語版 Wikipedia が実装されてからの，管理者の編集動向のデータ解析を行った．年々減少傾向になっている．

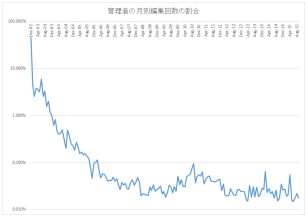

図 1　管理者の月別動向の割合

5. 考察

日本のインターネット文化の「完全な匿名性」の普及が大きな要因と考えられる．[1] 管理者になるとユーザー登録をすることになり，匿名での編集が出来ない．そのことが管理者の動向が減少していると考察する．

Wikipedia の編集履歴データを処理する際に，約 10 年分のデータを扱った．管理者の動向の調査だが「一般編集者から，途中から管理者になった編集者」の編集履歴の区別をすれば，更に正確なデータが取得できた．

6. 結論

日本語版 Wikipedia では，管理者の動向が品質に影響しているわけではなかった．管理者が不足しているにもかかわらず，プロジェクトが日々動いている要因として，日本の文化はどこも比較的均一なため，あまり用法や言動についての論争があまり生まれないことが挙げられる．

管理者の動向を調査するため，約 10 年分のデータを扱い解析し，ビッグデータを扱う処理技術を得た．

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　世界最大の百科事典はいかにして生まれたか. 株式会社早川書房, 2009.