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. 研究の背景

Wikipedia は，多くのボランティアにより，始まってから 10 年足らずの間に，大きな成長を見せたオンライン百科事典プロジェクトである．総記事数の文字数は 10 億文字を超え，ブリタニカ国際大百科事典とエンカルタ総合大百科の合計と比較しても上回る．Wikipedia は，さまざまな言語が参加しているグローバルなプロジェクトでもある [1]．2015 年 9 月までには，291 個もの言語が参加している．

このオープンなプロジェクトの百科事典は，制限なく誰でも自由に使用でき編集することもできる．

誰でも自由に編集できるからこそ，ボランティアの人々は気軽に参加でき，特定の企業や個人のお金を稼ぐのに力を貸していると感じることなく，時間と労力を注ぐことができる．

記事の内容はボランティアの人々の協力によって，信頼のおける品質が保たれている．しかし，中には協力的では無く，悪意のある編集をするものがいる．悪意のある編集者はその記事の内容とは関係ないことを書き込んだり，記事の破壊行為を繰り返している．Wikipedia では，悪意のある編集をする人とわかっていても規制などをしたりはしない．記事は完成・確定されることはなく，新しい情報にいつでも改変することができる．

本研究では，Wikipedia の全編集データをマイニングすることによって，Wikipedia の品質が保たれている理由を見つけ出す．

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品質管理がどのように行われているか調査する．この調査により，オープンな共同作業プロジェクトにおける，品質管理マネジメントの在り方についての知見を得たい．

3. 研究方法

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. どのような品質管理が行われているかその分析から調査する．
3. オープンなプロジェクトにおける品質管理マネジメントの在り方を提案する．

4. 成果物のイメージ

差し戻しに関するデータを収集し，編集回数や頻度などの要素を洗い出す．そして，いくつかの要素から条件を決めクラスター分析を行う．その結果から悪意のある編集がされている記事に共通する点を見つけ，Wikipedia のオープンなプロジェクトでの品質マネジメントの知見を得る．

5. 進捗状況

ビッグデータを解析するためのウェブサービス BigQuery で，Wikipedia のデータを提供されている差し戻しデータを抽出することができた．BigQuery が提供しているデータは、英語版のみでだった．他言語版を解析する為には，別の解析方法をとる必要がある [2].

6. 今後の計画

以下の順番で行う．

1. パソコンの環境をローカルで解析するために整える．
2. 日本語版の全履歴データを Wikimedia というサイトからダウンロードする．
3. Wikipedia の全履歴データを解析し，オープンなプロジェクトをする際の品質管理のあり方について調査し提案する．

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　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.

[2] Bigquery. https://cloud.google.com/bigquery/?hl=ja（2015.09.03 閲覧）.