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 億語を超え，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文字数は上回るほどある．2015 年 9 月には，Wikipedia の言語は 291 個も開設されており，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].

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金儲け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．また，顔や素性が分からない人たちと信頼し合い，共同作業で作り出されている．

しかし、編集者のすべてが善意を持っているとは限らず、中には悪意のある編集をするものもいる．悪意のある行為をする人とわかっていても Wikipedia では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．それにもかかわらず，我々が Wikipedia を使用している際はそのような記事は見かけず，信頼のおける品質が保たれている．

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成功理由を見つけ出す．

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

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い，どのような品質管理が行われているか調査する．また，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、英語版のみであり，他言語版を解析する為に別の解析方法をとる必要がある [2].

6. 今後の計画

1：ローカルで解析するためにパソコンの環境を整える．

2：Wikimedia というサイトから日本語版の全履歴データをダウンロードする．

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）.