千葉工業大学　社会システム科学部

プロジェクトマネジメント学科

平成27年度　卒業論文（中間）

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

矢吹研究室

1242005　石井康之

目次

**目次項目が見つかりません．**

第1章

序論

本章の構成

第1章では，本論文の序論を述べる．研究背景，研究目的，研究方法，プロジェクトマネジメントとの関連性，成果物のイメージ，本論文の構成について述べる．

研究背景

　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．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，制限無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．

研究目的

　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あり方についての知見を得たい．

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あり方を提案する．

プロジェクトマネジメントとの関連

　本研究は，プロジェクトマネジメントを学ぶことを目的としているため，プロジェクトマネジメントとの関連が全面的にある．

　考案するゲームは，プロジェクトマネジメント知識体系ガイド(PMBOK@ガイド)の第４版（以下，PMBOK）を参考にし，プロジェクトマネジメントとの一連の活動や，PMBOKに記載されている9つの知識エリアについての内容を活用する．

　PMBOKとは，プロジェクトマネジメントに関する知識体系である．

　現在は，PMBOKに従ってプロジェクトマネジメントを実施することが，デファクトスタンダードになっている．

　PMBOKに記載されている9つの知識エリアとは，何をやるべきかという観点，何を管理するべきかという観点からみたものである．

　9つの知識エリアは，以下9つのマネジメントについての内容となっている．

1. プロジェクト統合マネジメント
2. プロジェクト・スコープ・マネジメント
3. プロジェクト・タイム・マネジメント
4. プロジェクト・コスト・マネジメント
5. プロジェクト品質マネジメント
6. プロジェクト人的資源マネジメント
7. プロジェクト・コミュニケーション・マネジメント
8. プロジェクト・リスク・マネジメント
9. プロジェクト調達マネジメント

　本研究では，上記のPMBOKの中の品質マネジメントが関連性が最もあるものであるといえる．このオープンなプロジェクトの百科事典は記事の作成や，編集が主に行われて創り上げられており，その成果物がこの百科事典である．

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．

本論文の構成

　第1章では，序論を記述する．第2章では，本論文の研究目的である，Wikipediaについて，Wikipediaの記事や編集などについて，Wikipediaに記録されている編集履歴データがどのようなものかを調査する．第3章では，Wikipediaの編集履歴データを取得して，品質管理マネジメントがどのように行われているか調査する．第4章では，具体的な方法や過程を記述する．第5章では，本研究についての調査結果・考察，まとめを記述する．

第2章

Wikipediaについて

本章の構成

　本章では，本論文の研究対象であるWikipediaとは，Wikipediaの記事や編集など，Wikipediaの編集履歴データについて記述する．

Wikipediaとは

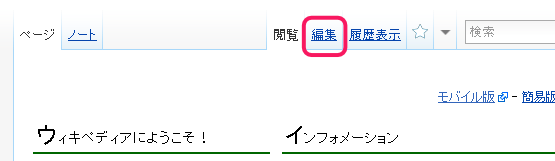
　フリー・ライセンスの百科事典である．フリーには2つの意味がある．無料という意味と，自由という意味だ．Wikipediaのフリーは後者の自由という意味であり，四つの自由が与えられている．著作物を複製する自由．改変する自由．再頒府する自由．そして，改変版を再頒府する自由だ．そして，営利目的に使っても，非営利に使ってもかまわない．というものがある．Wikipediaがフリーの百科事典であるというのは，無料でアクセスできるということではなくて，自由に複製，改変，利用してかまわないということである．

　Wikipediaという名前は，ウェブブラウザ上でウェブページを編集することができるWikiというシステムを使用した百科事典であることに由来する造語である．設立者の1人であるラリー・サンガーにより命名された．

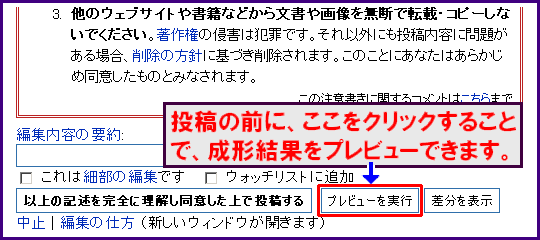
　Wikipediaは2015年9月までには，291個もの言語が参加している．この百科事典は多くの言語のボランティアたちによって書かれたグローバルなプロジェクトでもある．

記事の編集の仕方

　一部の保護されているページを除いて，全てのページには「編集」と書かれたリンクがあり，このリンクを使って，あなたが閲覧しているページを編集することができます．編集ができることはウィキペディアの大きな特徴で，この機能を使って，あなたが記事を修正したり，記事に加筆することができるのです．記事に情報を加筆する時には，情報の出典を明記してください．出典が不明な記述は，除去の対象となります．



これから常に使ってほしい大切な機能が「プレビューを表示」ボタンです．サンドボックスでなにか編集をして，それから「以上の記述を完全に理解し同意した上で投稿する」ボタンではなく，「プレビューを表示」ボタンを押してみましょう．そうすると，あなたがページに加えた変更の結果を，実際に保存する前に確認することができます．間違いは誰にでもあります．この機能は，間違いがないか自分で確認するためのものです．また，「プレビューを表示」ボタンを使えば，試しにページの体裁や表現をいろいろと変えてみても，ページの変更の記録にいちいち記録されずにすみますし，他にもいろいろと利点があるのです．でも，プレビューをした後，最後には保存するのを忘れないでください．



「以上の記述を完全に理解し同意した上で投稿する」ボタンを押す前に、あなたが行った編集の説明を、編集用のテキストボックスと保存ボタンの間にある要約欄に書き込むようにしましょう。ウィキペディアでは、ここに編集の説明を書き込むことが大切なエチケットと考えられています。ただ単に誤字を直したような時には「誤字修正」と書けば充分です。文章の意味に影響を及ぼさないような、小さな修正のときには、要約欄の下にある「これは細部の編集です（説明）」のチェックボックスにチェックをいれておいてください（この機能はログイン時にのみ有効です）。



第3章

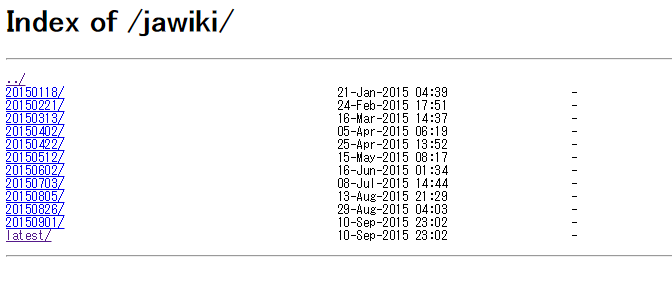
Wikipediaの編集履歴データの

取得について

Wikipediaの編集履歴データ

　データCreative Commons Attribution-ShareAlike 3.0 Unported License (CC-BY-SA) および GNU Free Documentation License (GFDL) の下にライセンスされており （Wikipedia:著作権および利用規約を参照），再配布や再利用のためにデータベース・データの提供が行われています．データの生成は不定期に行われている．

　Wikipediaではクロール行為のデータダウンロードは禁止されている．強引なクローリングは，Wikipediaが劇的に遅くなる原因となってしますためである．データベースから自動的にデータ収集している行為が発券された場合、システムの管理者から自身のサイトからWikipediaのアクセスを禁止されてしまう措置が起こってしまうこともある．また，ウィキペディア財団が法的措置を検討する場合もあるので，注意が必要．

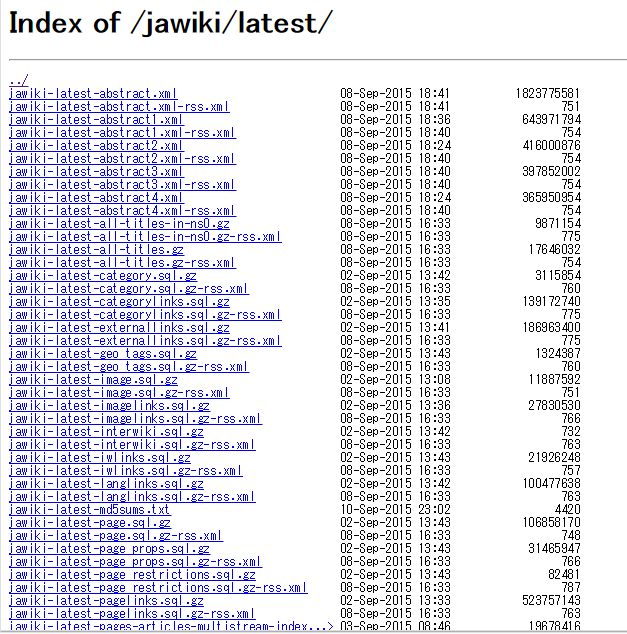


ここに日本語版Wikipediaの履歴データが記録されている．

URLはhttps://dumps.wikimedia.org/jawiki/

他の言語もこのような形式で履歴データが残されている．他の言語のデータを取得したい場合はURLのhttps://dumps.wikimedia.org/○○wiki/の○○の部分を変更すればよい．言語は英語のスペルで頭文字2文字でよい．

例：英語の場合はスペルはEnglishなので，https://dumps.wikimedia.org/enwiki/とすればよい．



どれか開くと上記のような画面になる．

ウィキページのデータはSQLのテーブルではなく、XMLで提供されている。XMLファイルの文字エンコーディングはUTF-8である。 非常にファイルサイズが大きいため、通常のエディタやブラウザで、解凍してはいけない。

データの詳細は下記のとおり

・pages-articles.xml.bz2 - ノートページ、利用者ページを除く最新版のダンプ

・pages-meta-current.xml.bz2 - 全ページの最新版のダンプ

・pages-meta-history.xml.7z - 全ページの全ての版のダンプ

・all-titles-in-ns0.gz - 全項目のページ名一覧 (標準名前空間)

Wikipedia:編集回数の多いページの一覧

期間: 2014-07-01 — 2014-07-31 のランキングです．

|  |  |  |  |
| --- | --- | --- | --- |
| 順位 | ページ | 編集回数 | 総編集回数 |
| 1 | 利用者:タペストリー/sandbox | 651 | 917 |
| 2 | Wikipedia:管理者伝言板/投稿ブロック/ソックパペット | 379 | 3098 |
| 3 | 利用者:ワーナー成増/sandbox | 376 | 2475 |
| 4 | Wikipedia:管理者伝言板/投稿ブロック/history20140727 | 368 | 4090 |
| 5 | Wikipedia:メインページ新着投票所/新しい項目候補 | 328 | 10594 |
| 5 | ハピネスチャージプリキュア! | 328 | 1878 |
| 7 | FNS27時間テレビ（2014年） | 306 | 306 |
| 8 | Wikipedia:改名提案/history20140727 | 283 | 6760 |
| 9 | 利用者:Tribot/log | 272 | 1720 |
| 9 | 利用者:ワーナー成増/下書き | 272 | 363 |
| 11 | 2014年のテレビ（日本） | 237 | 2321 |
| 12 | インテリビレッジの座敷童 | 229 | 233 |
| 13 | 洪門 | 209 | 243 |
| 14 | 義経＝ジンギスカン説 | 200 | 716 |
| 15 | 妖怪ウォッチ | 198 | 618 |
| 16 | スカッとゴルフ　パンヤ | 197 | 1514 |
| 17 | 笑福亭べ瓶 | 192 | 394 |
| 18 | 博士と助手～細かすぎて伝わらないモノマネ選手権～ | 178 | 1503 |
| 19 | マレーシア航空17便 | 168 | 168 |
| 20 | 小保方晴子 | 161 | 862 |
| 20 | Wikipedia:リダイレクトの削除依頼/2014年7月 | 161 | 161 |
| 22 | 大相撲力士一覧 | 151 | 1499 |
| 22 | 花子とアン | 151 | 1147 |
| 22 | 利用者:チンドレ・マンドレ/sandbox | 151 | 516 |
| 22 | ノート:集団的自衛権 | 151 | 280 |
| 26 | ALDNOAH,ZERO | 149 | 190 |
| 27 | ノート:野々村竜太郎 | 143 | 143 |
| 28 | 赤穂市 | 142 | 707 |
| 29 | STAP研究と騒動の経過 | 140 | 180 |
| 30 | Wikipedia:コメント依頼/みしまるもも 20140528 | 139 | 209 |
| 31 | GENEZ | 136 | 259 |
| 32 | ジェンパクト・ヘッドストロング・ビジネスコンサルティング | 134 | 134 |
| 33 | Wikipedia:保護依頼:history20140727 | 129 | 4587 |
| 34 | 静岡市 | 122 | 3252 |
| 35 | 利用者:ワーナー成増 | 119 | 419 |
| 36 | 利用者:ワーナー成増/下書き2 | 117 | 176 |
| 37 | 計報 2014年 | 116 | 877 |
| 38 | 利用者:Gowithitjam/sandbox | 115 | 115 |
| 39 | 仮面ライダー鎧武/ガイム | 113 | 2793 |
| 39 | 帝京大学 | 113 | 2743 |
| 41 | 2014 FIFAワールドカップ | 112 | 607 |
| 42 | ドラえもん（1979年のテレビアニメ）の帯番組時エピソード一覧 | 111 | 150 |
| 43 | パワーパフガールズ | 109 | 3655 |
| 43 | 利用者:Psyshotic Blue/下書き2 | 109 | 896 |
| 43 | 利用者:南北円上王 | 109 | 285 |
| 43 | Wikipedia- ノート:管理者への立候補 | 109 | 265 |
| 47 | 烈車戦隊トッキュウジャー | 108 | 958 |
| 48 | Wikipedia:コメント依頼/history20140727 | 106 | 3119 |
| 48 | 利用者:Tamrono157/サンドボックス | 106 | 144 |
| 50 | パナソニックショップ | 104 | 851 |
| 50 | 2014 FIFAワールドカップ・決勝トーナメント | 104 | 140 |
| 52 | 利用者-会話:Enyokoyama/sandbox | 101 | 251 |
| 53 | 金田一少年の事件簿（テレビドラマ） | 100 | 1208 |
| 53 | 東海中学校・高等学校 | 100 | 1028 |
| 53 | 刺激惹規性多能性獲得細胞 | 100 | 766 |
| 56 | さばげぶっ! | 99 | 171 |
| 56 | ドラえもん（1979年のテレビアニメ）のエピソード一覧（2001年 - 2005年） | 99 | 119 |
| 58 | Wikipedia:統合提案/history20140727 | 97 | 4536 |
| 59 | Wikipedia:削除の復帰以来 | 96 | 1833 |
| 59 | HERO（テレビドラマ） | 96 | 841 |
| 59 | RAIL WARS! -日本國裕鉄道公安隊- | 96 | 199 |
| 62 | 利用者:舎利弗/アンコール・ワット | 95 | 95 |
| 63 | SASUKE | 94 | 4100 |
| 64 | Wikipedia:議論が盛んなノート | 93 | 646 |
| 65 | 2014年の日本競馬 | 90 | 1063 |
| 66 | うえのやまさおり | 88 | 88 |
| 67 | 金田一少年の事件簿の犯罪者 | 86 | 907 |
| 67 | 利用者:Ajikoube-828/sandbox | 86 | 308 |
| 69 | 国際プロレス | 85 | 663 |
| 69 | GODZILLA ゴジラ | 85 | 525 |
| 69 | 実況パワフルプロ野球2013 | 85 | 488 |
| 69 | 利用者-会話:南北円上王 | 85 | 145 |
| 73 | 2014年のオールスター（日本プロ野球） | 83 | 83 |
| 74 | ハマトラ（アニメ） | 81 | 340 |
| 75 | 利用者:Quark Logo/sandbox3文禄・慶長の役 | 80 | 91 |
| 75 | 星亮一 | 80 | 80 |
| 75 | 俺の屍を超えて行け2 | 80 | 217 |
| 75 | 入江仁之 | 80 | 215 |
| 79 | 家族狩り | 79 | 132 |
| 79 | 山下達郎 | 79 | 2486 |
| 79 | ヘイトスピーチ | 79 | 686 |
| 79 | ノート:ゼーロン | 79 | 79 |
| 79 | ノート:橋本環奈 | 79 | 79 |
| 84 | 森川智之 | 78 | 2599 |
| 84 | 白雪姫 | 78 | 416 |
| 84 | バイナリーオプション | 78 | 89 |
| 87 | Wikipedia:分割提案 | 77 | 1645 |
| 88 | 橋本環奈 | 76 | 161 |
| 88 | 利用者:やまさきなつこ/sandbox | 76 | 128 |
| 88 | 牛丸謙壱 | 76 | 76 |
| 91 | DDTプロレスリング | 75 | 1596 |
| 91 | 利用者:K s/sandbox | 75 | 1375 |
| 91 | 利用者:Iso10970/sandbox | 75 | 205 |
| 94 | ウルトラマンギンガS | 74 | 112 |
| 95 | ガールズ&パンツァー | 73 | 1271 |
| 95 | チェルシーFC | 73 | 1176 |
| 95 | まじもじるるも | 73 | 140 |
| 98 | Wikipedia:Bot作業依頼 | 72 | 2161 |
| 98 | 札幌競馬場 | 72 | 688 |
| 100 | 浪曲 | 71 | 731 |
| 100 | 集団的自衛権 | 71 | 131 |
| 100 | 利用者-会話:B side of the moon | 71 | 112 |

第4章

手法

本章の構成

本章では，Wikipediaの履歴データを扱う方法を述べる．

研究を行うための用意

開発環境としてLinuxを扱う．そのために，CentOSとVirtualBoxを用意する．

VirtualBoxのインストール

1. ダウンロード

Oracleが提供しているOracleVitualBoxというのを下記のサイトからダウンロードする．本研究ではWindowsOSを使用しているので，VirtualBox platform packagesの中にある「VirtualBox 5.0.4 for Windows hosts」 というのを選択する．



↓ダウンロードサイト

https://www.virtualbox.org/wiki/Downloads

２．インストーラーの実行

下記のサイトからVirtualBoxの導入からOSインストールまでの解説が載ってるので参考にする．



↓参考サイト

http://success.tracpath.com/blog/2013/10/15/virtualbox%E3%81%AE%E5%B0%8E%E5%85%A5%E3%81%8B%E3%82%89os%E3%82%A4%E3%83%B3%E3%82%B9%E3%83%88%E3%83%BC%E3%83%AB%E3%81%BE%E3%81%A7%E3%81%AE%E3%83%81%E3%83%A5%E3%83%BC%E3%83%88%E3%83%AA%E3%82%A2/

Linuxのインストール

下記のサイトに「Linux標準教科書」というPDF形式のファイルがあるのでそれをダウンロードする．そのファイルの中の「第2章　Linuxのインストール」を参考にし，用意する．



↓ダウンロードサイト

http://www.lpi.or.jp/linuxtext/text.shtml

Wikipediaのデータから読み取ったデータと読み取り方について述べる．

2015年4月1日～2015年4月30日までに編集回数された回数が多い記事

