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

謝辞

目次

第 1 章

序論

第 2 章

背景

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．

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．

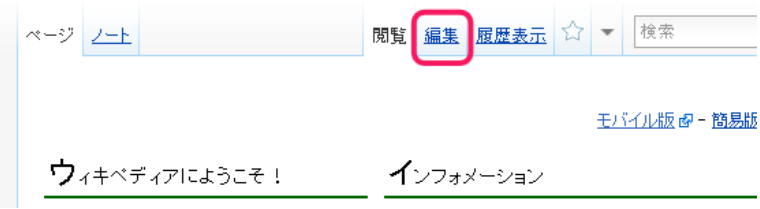


図 2.1　図の挿入例

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．

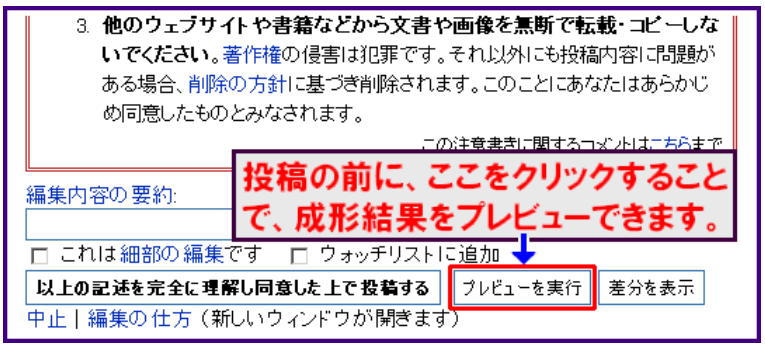


図 2.2　図の挿入例

「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）。

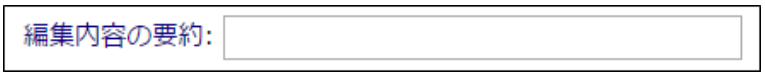


図 2.3　図の挿入例

Wikipedia の編集履歴データ

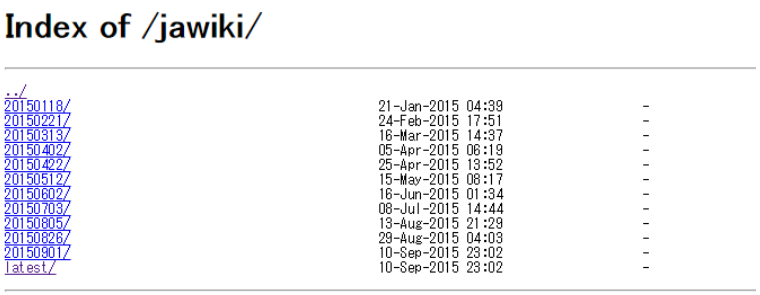
データ Creative Commons Attribution-ShareAlike 3.0 Unported License (CC-BY-SA) および GNU Free Documentation License (GFDL) の下にライセンスされており （Wikipedia:著作権および利用規約を参照），再配布や再利用のためにデータベース・データの提供が行われています．データの生成は不定期に行われている．

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うためである．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



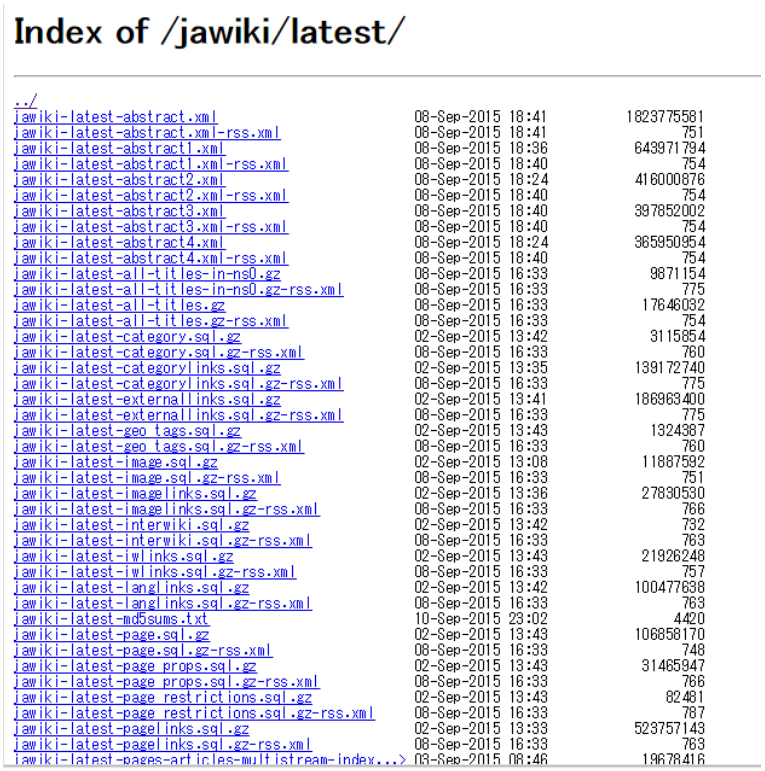
Index of /jawiki/

../		
20150118/	21-Jan-2015 04:39	-
20150217/	24-Feb-2015 17:51	-
20150313/	16-Mar-2015 14:37	-
20150402/	05-Apr-2015 06:19	-
20150422/	25-Apr-2015 13:52	-
20150512/	15-May-2015 08:17	-
20150602/	16-Jun-2015 01:34	-
20150703/	08-Jul-2015 14:44	-
20150805/	13-Aug-2015 21:29	-
20150826/	29-Aug-2015 04:03	-
20150901/	10-Sep-2015 23:02	-
latest/	10-Sep-2015 23:02	-

図 2.4　図の挿入例

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.



Index of /jawiki/latest/

../		
jawiki-latest-abstract.xml	08-Sep-2015 18:41	1823775581
jawiki-latest-abstract.xml-rss.xml	08-Sep-2015 18:41	751
jawiki-latest-abstract1.xml	08-Sep-2015 18:36	643971794
jawiki-latest-abstract1.xml-rss.xml	08-Sep-2015 18:40	754
jawiki-latest-abstract2.xml	08-Sep-2015 18:24	416000876
jawiki-latest-abstract2.xml-rss.xml	08-Sep-2015 18:40	754
jawiki-latest-abstract3.xml	08-Sep-2015 18:40	397852002
jawiki-latest-abstract3.xml-rss.xml	08-Sep-2015 18:40	754
jawiki-latest-abstract4.xml	08-Sep-2015 18:24	365950954
jawiki-latest-abstract4.xml-rss.xml	08-Sep-2015 18:40	754
jawiki-latest-all-titles-in-ns0.gz	08-Sep-2015 16:33	9871154
jawiki-latest-all-titles-in-ns0.gz-rss.xml	08-Sep-2015 16:33	775
jawiki-latest-all-titles.gz	08-Sep-2015 16:33	17646032
jawiki-latest-all-titles.gz-rss.xml	08-Sep-2015 16:33	754
jawiki-latest-category.sql.gz	02-Sep-2015 13:42	3115854
jawiki-latest-category.sql.gz-rss.xml	08-Sep-2015 16:33	760
jawiki-latest-categorylinks.sql.gz	02-Sep-2015 13:35	139172740
jawiki-latest-categorylinks.sql.gz-rss.xml	08-Sep-2015 16:33	775
jawiki-latest-externallinks.sql.gz	02-Sep-2015 13:41	186963400
jawiki-latest-externallinks.sql.gz-rss.xml	08-Sep-2015 16:33	775
jawiki-latest-geo_tags.sql.gz	02-Sep-2015 13:43	1324387
jawiki-latest-geo_tags.sql.gz-rss.xml	08-Sep-2015 16:33	760
jawiki-latest-image.sql.gz	02-Sep-2015 13:08	11887592
jawiki-latest-image.sql.gz-rss.xml	08-Sep-2015 16:33	751
jawiki-latest-imagelinks.sql.gz	02-Sep-2015 13:36	27830530
jawiki-latest-imagelinks.sql.gz-rss.xml	08-Sep-2015 16:33	766
jawiki-latest-interwiki.sql.gz	02-Sep-2015 13:42	732
jawiki-latest-interwiki.sql.gz-rss.xml	08-Sep-2015 16:33	763
jawiki-latest-iwlinks.sql.gz	02-Sep-2015 13:43	21926248
jawiki-latest-iwlinks.sql.gz-rss.xml	08-Sep-2015 16:33	757
jawiki-latest-langlinks.sql.gz	02-Sep-2015 13:42	100477638
jawiki-latest-langlinks.sql.gz-rss.xml	08-Sep-2015 16:33	763
jawiki-latest-md5sums.txt	10-Sep-2015 23:02	4420
jawiki-latest-page.sql.gz	02-Sep-2015 13:43	106858170
jawiki-latest-page.sql.gz-rss.xml	08-Sep-2015 16:33	748
jawiki-latest-page_props.sql.gz	02-Sep-2015 13:43	31465947
jawiki-latest-page_props.sql.gz-rss.xml	08-Sep-2015 16:33	766
jawiki-latest-page_restrictions.sql.gz	02-Sep-2015 13:43	82481
jawiki-latest-page_restrictions.sql.gz-rss.xml	08-Sep-2015 16:33	787
jawiki-latest-pagelinks.sql.gz	02-Sep-2015 13:33	523757143
jawiki-latest-pagelinks.sql.gz-rss.xml	08-Sep-2015 16:33	763
jawiki-latest-pages-articles-multistream-index...>	03-Sep-2015 08:46	19678416

図 2.5　図の挿入例

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.

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

図 2.6　図の挿入例

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

図 2.7　図の挿入例

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

図 2.8　図の挿入例

第 3 章

目的

第 4 章

手法

第 5 章

結果

第 6 章

考察

第 7 章

結論

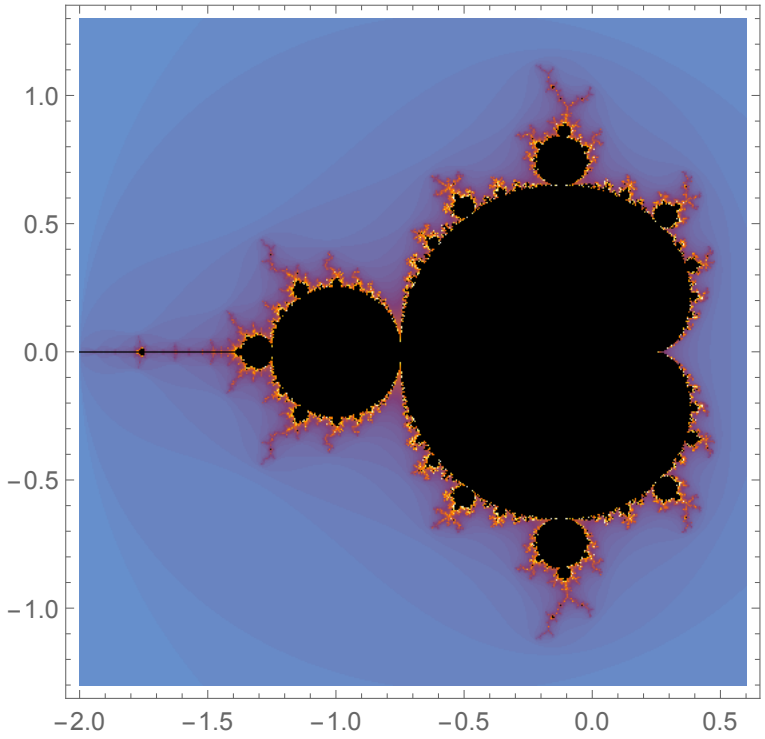


図 7.1　図の挿入例

参考文献は文献ファイル（この文書では `biblio.bib`）に記述し，`\cite` で参照する．例：データベースのための問い合わせ言語 SQL で数独を解く方法が提案されている [1]．このように参照すると，参考文献リストに自動的に登録される．文献の種類には，雑誌論文 [1] や会議録論文 [2]，卒業論文 [3]，書籍 [4]，ウェブサイト [5] などがある．文献の種類によって必要な項目が異なるため，`biblio.bib` を見て確認すること．

参考文献

[1] 矢吹太朗, 佐久田博司. SQL による数独の解法とクエリオプティマイザの有効性. 日本データベース学会論文誌, Vol. 9, No. 2, pp. 13–18, 2010.

[2] 矢吹太朗, 増永良文, 森田武史, 石田博之. 知識体系のエリア自動抽出のためのユニット分類手法. 第 5 回データ工学と情報マネジメントに関するフォーラム（DEIM2013）. 電子情報通信学会データ工学研究専門委員会, 日本データベース学会, 情報処理学会データベースシステム研究会, 2013.

[3] 久保孝樹. チケットを活用するオープンソースソフトウェア開発の実態調査. 卒業論文, 千葉工業大学, 2014.

[4] 奥村晴彦, 黒木裕介. LaTeX2e 美文書作成入門. 技術評論社, 第 6 版, 2013.

[5] 矢吹太朗. 自分のコードを出力するプログラム. http://www.unfindable.net/article/self.html（2012.12.01 閲覧）.