

ウィキペディアのデータベース解析と考察

永江 孝規

メディアアート表現学科

Analysis and Study on the Wikipedia Database

NAGAE Takanori

Department of Media Art

(Received November 6, 2009 ; Accepted January 14, 2010)

1. はじめに

公開されたオンラインのデータベースはそれ自体が興味深い研究対象である。しかも、それが実時間で、不特定多数の人々によって不断更新されているとすればなおさらではなかろうか。多くの場合それらは掲示板であったりコミュニティサイトであったりする。社会情勢にあわせて次々に新しいトピックが作られては蓄積されていく。またユーザたちがトピックからトピックへと関わりあうことによってトピック間またはユーザ間に複雑な関係性が生まれていく。このような社会現象を動的データベースの自己組織化というテーマで捉えたとき、ひとつの研究テーマとして成立し得る。従来手法の改良というのではなく、今まさに現在進行中の現象を、同時並行して扱って行くことになる。いわば、学問として評価基準が定まっていない分野を暗中模索しながら開拓していくようなものであると言える。

ウィキペディアは、我々研究者や教育者にとって、わずか数年の間に無視できない存在に成長した。いまや、ウィキペディアなしの研究や教育など考えられないだろう。しかし学術世界に属する人々の多くは、この新しい潮流に対して、まだうまく自分の立ち位置を確立できていないのではないかと[assouline 08]。

学術研究は、これまで実名で行うものであった。論文はもちろん、百科事典の執筆などもその類である。しかしウィキペディアの編集は通常、実名ではなく架空の「ユーザ」として行われ、固定の「ユーザ名」さえ持たず、「匿名ユーザ」として執筆に参加することもできる。しかもはなはだ気味の悪いことにこれらの共同編集者たちには多くの中学生や高校生ら、あるいはまったくの素人が含まれていると予想できる。かかる連中と半ば名前を伏せて共同で執筆作業をするなどということは、研究の専門職である大学教員らにとってはおよそ想定されざ

ることであろう。

だが多くの研究者にとってウィキペディアの魅力は抗し難いものがある。そこで、ある程度良心的な研究者であれば、学術研究というものは従来ギブアンドテイクなものであるから、ウィキペディアによって自分が得た恩恵以上の貢献を、ウィキペディアに返したいと考えるかもしれない。だが自分の名前を伏せたまま執筆に参加するのは何か損をしているような、釈然としないものを感じるかもしれないし、かといって実名でウィキペディアに書き込むことも躊躇されるだろうし、また仮に実名で執筆したとして、自分の書いたものを誰だかわからない輩に簡単に差し戻されたり批判されるのは叶わないだろう。察するに、研究者とウィキペディアのジレンマとはこうしたものではなかろうか。

本稿の執筆動機は、まず、このようにウィキペディアに対して研究者はどのように向き合うべきか、どのように付き合い、利用し、貢献していくべきか、についての自問自答であった。さらには、技術者として純粋に技術的興味から、ウィキペディアの巨大なデータベースやソフトウェア群を使いこなして、何か新しいWebサービスを作り出せないかと思った。やがて、いわゆる「推薦システム」「推薦エンジン」(recommender system, recommendation engine) と呼ばれる手法をウィキペディアデータベースに応用してみて、そこから得られるさまざまな「新しい結果」に魅了されていった。

ウィキペディアといういくつかの、それまでは別々に扱われてきた研究分野にまたがる題材を扱うにあたって、本稿はやや難解な構成を取らざるを得なかった。本稿を極力シンプルに保つために、本稿のメインテーマであり提案手法でもある「推薦システムのウィキペディアデータベース解析への応用」に限った構成にしようかとも思った。しかし筆者は、本稿を単なる技術論文にしたいとはなかったし、そもそも単なる技術研究を行いたいと思って

いるわけでもない。技術的な足がかりからさまざまな文化・芸術的话题へたどっていきけるからこそ、ウィキペディアは研究のしがいがあるからである。特にウィキペディア日本語版に関してはそのメインカルチャーに対するサブカルチャーの優越性に極めて注目すべき特徴がある。筆者は必ずしもサブカルチャーの専門家ではないが、ウィキペディア日本語版と関わりをもつ範囲で、筆者の現時点の考察を本稿に盛り込んでおきたかった。

そこで「推薦システムを適用した提案手法」について本稿の第4章において、いわば隔離して述べることにした。技術論的な新規性の評価などはこの章を参照していただきたい。この提案手法とは現在 Web で公開している手法[wpw]という意味であり、また「推薦システム」とは関係ないが従来手法にはない手法についてもここで紹介している。

さかのぼって第2章では本稿を構成する上で必要最小限なウィキペディアについての概要について述べる。必要最小限というのはつまり、ウィキペディアについてあれこれと言及しだすとあまりにも多くの話題が交錯し収拾つかなくなるので、いわば意図的にその論述する領域を制限しているのである。概要と言ってもそれは世間一般に言われているウィキペディアの全体像の俯瞰というよりも、筆者の関心事に偏った、ある意味では本稿独自の指摘あるいは考察と言った内容になっているかもしれない。また、やや紛らわしいが、筆者の提案手法ではない従来手法を用いて（つまりごくオーソドックスな統計手法によって、という意味だが）筆者が独自に解析した結果も既存手法[Zeit]とともにここで述べている。以上の事柄は技術的な新規性とは直接関係がないので第4章と別立てで論じていると考えていただきたい。

第3章では解析に必要な技術的準備、従来手法の紹介などについて述べている。

第5章は、本稿全体にわたる総論、まとめ、今後の課題などであるが、お読みいただければそれが単なる技術論ではないことがおわかりになると思う。ただこれが何論と呼ぶべきものであるか筆者ですら今の段階ではよくわかっていない。

本章、第1章においては簡単な導入、本稿の執筆動機と、本稿の構成について述べた。くどいようだが、第3、4章は技術論であって、そのうち第4章の方に本稿の新規性が盛り込まれている。第2、5章にも筆者の考察は含まれているが、それらを第3、4章とは混在させたくなかったのでこのような論文構成となった。

2. ウィキペディアとは何か

2.1 概要

ウィキペディアの総元締めはウィキメディア財団という米国に設立された法人（フロリダ州法による米国内国歳入法501条(c)(3)認定を受けた慈善団体）であるが、その設立（2003年6月20日）はウィキペディア日本語版開設（2002年5月）よりも遅い。ウィキペディアの種々のサービスが展開していく過程で発生した諸問題を支援するために後追いで作られた組織であると言える。ウィキメディア財団の他に国などの地域ごとに財団から公式に認められた地方支部（local chapter、wikimedia chapter）という法人格をもった組織があるが、現時点（本稿以下で「今」「現在」「現時点」とは2009年11月を意味する）でそれらを承認された日付の古い順に列挙すると次のようになる[chap]。

1. ドイツ（Deutschland 2004/06/13）
2. フランス（France 2004/10/23）
3. イタリア（Italia 2005/07/17）
4. ポーランド（Polska 2005/08/14）
5. セルビア（Serbia 2005/12/03）
6. スイス（CH 2006/05/14）
7. オランダ（Nederland 2006/03/27）
8. イスラエル（Israel 2007/03/06）
9. 台湾（Taiwan 2007/07/04）
10. アルゼンチン（Argentina 2007/12/11）
11. スウェーデン（Sverige 2007/12/11）
12. オーストリア（Österreich 2008/02/26）
13. オーストラリア（Australia 2008/03/04）
14. 香港（Hong Kong 2008/03/04）
15. チェコ（Czech Republic 2008/03/06）
16. ロシア（Russia 2008/05/24）
17. ブラジル（Brasil 2008/10/07）
18. インドネシア（Indonesia 2008/10/07）
19. ノルウェー（Norge 2008/11/03）
20. ハンガリー（Magyarország 2008/11/03）
21. ニューヨーク市（New York City 2009/01/12）
22. 英国（UK 2009/01/12）
23. ポルトガル（Portugal 2009/07/03）
24. ウクライナ（Ukraine 2009/07/03）
25. デンマーク（Danmark 2009/07/03）
26. マケドニア（Macedonia 2009/09/21）
27. フィンランド（Suomi 2009/09/21）

ウィキペディア日本語版は、その利用者数は英語版に

次ぎ、記事数でも5位に位置するが、日本法人は存在せず、また設立される予定すらない。

ウィキペディアは、上記法人が設置された「国」や「地域」ごとではなく、「言語」ごとに別々のサイトで運営されている。世界中に何百という言語があり、自然言語もあれば人工言語や架空の言語もあり、わずか数千人しか話者がいないマイナーな言語もある（たとえば人工言語のエスペラントには native speaker が200人から2000人居ると言われている）。それらの言語のほとんどすべてでウィキペディアが運営されているか、運営がかつては試みられたか、試みられつつある（たとえば架空言語のクリンゴン語版は2008年3月に閉鎖され、Wikia [wikia] に移された [klngn]。漢字交じり韓国語版は Wikia には存在する [hanja] が、2008年5月にウィキメディアへの申請が否決された [reqhanja]）。

法人は通常、国ごとに置かれるものであるが、ウィキペディアは言語ごとに管理されている。しかし、言語に法人を置くことは不可能であって、このことがウィキペディアにおいてさまざまな問題の種となっている。また、言語と国がほぼ一致している日本のような場合にも、未だ法人が置かれていないという別種の問題もある。地域ごとではなく言語ごとに運営されていると言っても実際には地域ごとの特徴が出てくる。たとえば中国語版であれば、実質的には台湾が香港のユーザが編集している。編集回数上位の記事を調べてみれば明らかだが、台湾・香港に関する記事（「国立高雄第一科技大學」、「中華民国」、「翡翠台電視劇集列表」、など）の方が、中華人民共和国に関する記事よりもずっと多い。台湾、香港、中国の三地域が中国語版を共同運営していくとして、誰がその管理をとりまとめるのか、などということは誰にも想像がつかないだろう（中国語とはいわゆる北京語（普通話）のことであり、方言版「粵語（広東）」「呉語（上海）」などもあるが現時点で記事数は極めて少ない）。また、英語版はあってもカナダ版やオーストラリア版などは存在しない。カナダ人やオーストラリア人たちは英語版を共有している。将来英語版の多数派はインド人となって、英語版は事実上インド人とパキスタン人の百科事典になるかもしれない。スペイン語版も同様であって、スペイン版というよりは人口比率的には中南米版と考えた方がよい。

日本語版に関して言えば、管理者は皆、通常の登録ユーザの中から自薦・他薦によって選ばれた有志のボランティアであり、ウィキメディア財団に所属する者や雇用されたものが管理運営しているのではない。また、goo のような大手企業（つまり NTT グループ）がウィキペディアのデータベースを二次利用してはいるが [goo]、それ

はあくまでもウィキペディアがその記載情報をクリエイティブ・コモンズのライセンス（かつては GFDL: GNU 自由文書ライセンス）で公開しているからであり、goo とウィキペディアの間でなんらかの契約関係が結ばれていることを意味していない。また、goo もウィキペディア由来の情報について一切責任を負っていないし、またウィキペディアとの内容の同一性も保証していない。

このように、ウィキペディアは、極めて曖昧模糊とした形で運営されている。こうした組織の特徴として、不特定多数の人たちが集まってきて、その中から何人かが管理を任されて（逆に管理者を解任されたりして）、さまざまなトラブルに巻き込まれながら徐々に役割分担や対処方法が確立していき、長い試行期間と合意形成を経て（とは言ってもせいぜい10年くらいだが）、今のところ表面的には無事運用されているように見える。ネット検索では常に上位に位置しており、認知度も非常に高い。しかし、良く調べれば調べるほどその立脚している基盤や根拠は弱く不明瞭に思えてくる。

たとえば、ウィキペディアのパロディサイトであるアンサイクロペディア [Unc] では、ウィキペディアを「宣伝サイト」「オタクのオタクによるオタクのための百科事典」と称しており、一種のジョークの形を取っているが、それらのことはある程度まで事実である。また、「2ちゃんねるの百科事典」「google の百科事典」、あるいは「電通の百科事典」などと揶揄しているが、これらの言及にもそれなりに深い意味合いが込められていると言って良いだろう。

2.2 ウィキペディアと MediaWiki

ウィキペディア日本語版の過渡期は今日から見れば2006年から2007年くらいであり、この頃に主要な問題の多くが噴出し [時実07]、また管理体制にもかなりの進展が見られた。「ウィキペディアで何が起きているか」という書籍は、2008年中頃に出版され、それまでの顛末の多くが記録されている [山本・古田08]。同時代の貴重な資料としては他に池田信夫ブログ [池田07] を加えればほぼ網羅できるだろう。過去の熾烈な編集合戦の歴史についてはこれらの文献を参考にしていただきたい（ウィキペディア関連書籍は実はこの時期にまとめて出版されたが、それから二年が経ち、状況はずいぶん変わってきている。しかしその新しい動向について本稿では特に必要ない限り言及していない。また可能な限り最新の状況に基づいて記述している）。

当時は、ウィキペディアに代わる第二、第三の共同執筆オンライン百科事典が現れていた。ウィキペディアの創設者の一人が立ち上げた Citizendium [Cit]、ウィキペ

ディア日本語版の管理体制と対立して独立した Yourpedia [You]、google が提案した knol [Knol] などである。Yourpedia はウィキペディアとまったく同一のソフトウェア MediaWiki [Medw] によって動いているが、それは、ウィキメディア財団が、そのソフトウェアごと無償で公開しているからである。MediaWiki を用いてウィキペディアと同様のウィキサービスを提供しているサイトとしてはさきほど挙げたアンサイクロペディアなど多数ある。また、ウィキメディア財団が提供する類似のウィキプロジェクト (Wiktionary [Wikt]、WikiNews [Wikn] など) もあるし、ウィキペディアの創設者が関わっているホスティングサービス Wikia [Wikia] などもある。Wikia に至っては、MediaWiki のインストールもサイト管理も不要で、だれもが MediaWiki の管理者になることができる。従って、ウィキペディアの運営に不満があれば、誰でもウィキペディアに取って代わることは可能な状況にある。

これらの多くの試みはもしかするとウィキペディアを凌駕するかもしれないと期待され、ウィキペディアはその管理体制の脆弱さから、ユーザに見捨てられていずれ廃れるのではないかと、とも言われた。しかしながら現状を観察してほぼ確実に言えることは、普及度やユーザ数、記事数などのどれをとってみても、ウィキペディアより優れたウィキサービスは存在しない、ということだ。今や、ウィキペディアがあらゆるウィキの代名詞となった感がある。

だがウィキペディア以外の選択肢が見あたらないということは、ウィキペディアの問題がすでに解決済みであることを意味しないし、今後もウィキペディアが発展しつづけることを保証もしない。ウィキペディアの発展や可能性についていろいろと考察を述べたい気持ちもあるが、本稿のテーマからややはずれるために割愛せざるを得ない [Ortega 09] [WSJ 09] [Peel 09]。

ウィキペディア日本語版の運営は、現地法人化されないうまま、なんとなく安定してしまった。非常に気持ちが悪い、中途半端な状態だ。たとえるならば、大学教員など有志が集まって学会を立ち上げたが、人的リソースや金銭的事情で、組織も事務所もあるようでないような状態である、と言える。そのような同好会レベルの NPO はたくさんあるだろう。台湾や香港などのように比較的小規模な地域の方が現地法人をある意味「気楽」に立ち上げられるという事情もあるのかもしれないが、一般ユーザとほとんど同程度の責任しか負えない管理者に負担がこれ以上かからないように、日本にも法人格を持つ支部くらいできてほしいというのが個人的な希望である。

なお、筆者らが本学メディアアート表現学科のウィキ

[Meda] を MediaWiki で立ち上げたのは 2006 年 4 月 1 日であったが、このときにはまだ MediaWiki は当時数多く知られていたウィキソフトの一つに過ぎず、また必ずしもよく知られたソフトでもなかった。当時日本で最も利用されていたのは PukiWiki [Puk] というもので、筆者らも 2005 年頃まではこの PukiWiki を使っていた。PukiWiki などのデータベースを使わないウィキソフトは記事の一つ一つローカルなファイルとして保存するので、記事数が数千程度までの小中規模のウィキの場合、問題なく快適に動作したので、当時の「まとめサイト」には PukiWiki がよく使われていた。しかし数万、数十万の記事ができると非常に動作が重くなる。OS がそんなに大量のファイルを管理できなくなるからである。筆者らは実験の都合上、PukiWiki に代わるより強力なウィキソフトを探していて、MediaWiki に巡りあった。

記事は個別のファイルではなくデータベースとして管理すれば、安定して且つ高速にアクセスできる。そこで、MySQL サーバなどを使った MediaWiki などのウィキソフトが徐々に台頭してくる。この傾向はブログなど他のウェブサービスでも同時進行的に進んだ。

2.3 リビジョン管理

PukiWiki などのやや古い世代のウィキシステムにはリビジョン管理機能がないが、MediaWiki にはこの機能が最初から備わっている。最近では Wordpress などのブログもリビジョン管理ができるようになった。

ウィキペディアにおけるリビジョン管理について、最初に少しか説明しておく。必要があれば、そのつど詳しく再説することにする。

ユーザの誰かが新しく項目を執筆したとき、それを初版と言う。既存の項目に 1 字でも修正・加筆を行えば、その項目は新しい版 (リビジョン) となる。このような時系列の一連の版の集まりを編集履歴という。編集とは版を更新し、履歴を保存していくことに他ならない。

通常すべての版はそのまま無期限に保存される。ただし、著作権侵害や個人情報を含む版はユーザや管理者らの合議を経てまるごと完全に消去されて履歴にも残らない。これを特定版削除という [特定版削除]。しばしば起こることだが、初版を未熟な匿名ユーザが執筆して、しかもそれがどこかのサイトの丸写しだったりすると、初版以後すべてが著作権侵害となって、項目自体が履歴を含めてすべて削除されることがある [著作権侵害]。

編集履歴があまりにも長くなると、技術的な理由で古い履歴が別の項目 (ログ) として分離されることがある。このことをログ落ちと言うことがある。このようにログとして分離されると同じ項目の履歴として古い版を参照

できないことがあるが、ログを探すことによって初版までたどることができる。最新版から何か文言を削除しても、特定版削除でない限り、それらの文言はいつまでも履歴に残り続けるということに注意していただきたい。

2.4 編集履歴解析の歴史的意義

西欧では何世紀もの歳月をかけて学問（特に「神学」）の一分野として確立された「聖書学」というものがある。聖書の写本の間の変異を比較することによって、誤りや改竄を発見し、本来の原典の文言を復旧しようというもので、その手法によって「本文批評」「高等批評」などに分類される[Mack 05][加藤99][Ehrman 06]。新約聖書に限って言うならば、歴史的に最も早く成立したのはパウロ書簡の中の「テサロニケの信徒への手紙1」であり、西暦50年頃に書かれたとされ、福音書では「マルコの福音書」が西暦70年頃に書かれたと言う。いずれも、教祖イエスの死後、20～40年を経た後に書かれていて、かつ、イエス本人から直接取材したのではない。その他の福音書はさらに後になって、さまざまな史料から編集されたものであることがわかっているし、最古の写本は4世紀以降のものしか残っていない（シナイ写本、ヴァチカン写本など。ただし写本の断片はもっと古くさかのぼれると言う）。そもそも、今日の我々の感覚からしても、ある人が死んで、40年後に初めて伝記として文書化されたとして、それにどれほどの信憑性があるだろうか。たとえば1945年に何か事件があったとして、その間読み書きのできる人、リテラシーのある人がまったくおらず、すべて口伝で伝わってきたとして、それを1985年になって初めて記録したとする。それはどの程度正確だと言えるだろうか。

我が国でも「平家物語」は50年ほどの歳月をかけて、多くの人たちによる改訂を経て完成したものと言う。そして義経の鶴越がほんとうの史実であるかどうかを確認することは絶望的に不可能だ。戦闘の記録というものは、その場ですぐに記録しても曖昧であるのに（たとえばノモンハン事件や盧溝橋事件、南京事件など。サイパン戦や沖縄戦ですらそうだ）、50年後に民間伝承だけで確かな事実を知ることなど不可能だろう。「平家物語」の影響を受けつつ成立した「太平記」では、比較的実時間に記述されていたと言われるが、結局のところ、新田義貞がどのルートからどうやって鎌倉に侵入したのかとか、新田義貞と足利義詮（千寿王）のどちらが主導権を握っていたのかなど、肝心なことは定かでない。こうした中世の軍記物などは誰がいつ書いたのか、それを後世の人がどのように編集し改竄あるいは捏造したのかなど、まるでわからないのが普通である。

一方でウィキペディアは、そのプロジェクトが始まった最初から、すべて誰がいつどの箇所を執筆し、編集したかという記録が完全に残っている（少なくとも日本語版が公開された頃までにはそのような仕様になっていた）。たとえば、データベース中に現存する日本語版の最古のページは「エスペラント」「欧州連合」「国の一覧」などである。これらは2002年9月10日に現れている。ウィキペディアが日本語に対応したのは2002年9月1日であった。また、「エスペラント」「欧州連合」などの項目は古い別のウィキからの転載であったらしい。またこれらを書き込んだのは Brion VIBBER というアメリカ人（日本語版の最初の管理者）であり、英語、エスペラント語、フランス語、スウェーデン語などを話し、また日本語の読み書きも若干できるユーザであることがわかっている。このように、過去の記録に遡ってみて初めて、ウィキペディアプロジェクトというものが、言語学者、特にエスペランティストらによって創始されたということがわかってくる。

ウィキメディア財団はこれら過去の完全な編集履歴をダンプデータとして、無償で提供している。何か不審な記事を見つけたときは、このダンプデータを解析することによって、心ゆくまで誰によるどんな編集であるかを調べ上げることができる。つまり聖書学的な本文批評を、「仮説」や「推測」による補完無しで、完全な形で適用できるということである。

完全な編集履歴の保存という発想はおそらくはプログラミング開発の世界で始まったことだろう。多くの人が共同でプログラムを開発するにはソースコードのリビジョン管理が必要になってくる。特にフリーソフトウェアやオープンソースのソフトウェア開発は、世界中のさまざまな人たちが関わり合ってくる。ある閉じた企業の開発室よりは、このような世界でまず、リビジョン管理は必要とされたのではないか。このようなリビジョン管理という考え方は、今日では Microsoft Word や OpenOffice.org Writer などの普通のワープロの校正にも使われるようになってきた。聖書や平家物語が執筆された時代から、今日のリビジョン管理された著述の時代への変遷というものは、歴史的必然であって、その最先端で最大規模の典型がウィキペディアであると言って良い。

ほんの数十年前まで、我々は、手書きで原稿用紙に向かって執筆していた。かつての百科事典は、出版社があり、編集長と編集者と執筆陣が居て、編集会議を行って、何度も校正してできあがったものである。電子機器がなかった時代はそうやって書籍を作るしかなかったし、彼らなりに正確を期して作業したに違いない。しかし、我々はその編集履歴をもはや永久に知ることはできない。か

つては原稿以外のメモ書きや草稿などは（夏目漱石や芥川龍之介レベルの著名人でなければ）世の中に出ることがなかった。おそらく執筆者本人がそんなことまで覚えていないし残してもいなかったらう。その執筆や編集の過程を知らずとも、我々は小説を楽しみ、事典を利用することはできる。同じことはウィキペディアについても言えるのであり、我々は別にウィキペディアがどんな仕組みでできていて、誰が執筆しているかなど知らなくても利用することはできる。ウィキペディアから恩恵をうける99.9%までの人たちにとって、ウィキペディアの編集履歴がどんなものかを知る必要はない。しかし、ウィキペディアの最大の特徴は、少なくともその一つは、編集履歴が完全に保持されていることである。

最終的に紙に記録されることがないデジタルメディアは、簡単に修正される。このことについて、アマゾン・コムのキンドルやニコニコ動画のコンテンツ削除問題を例に挙げて、山形浩生氏が問題提起をした[山形09]が、デジタルメディアは同時に厳格なリビジョン管理もできる媒体であって、かつその意志さえあれば、紙などよりもずっと長期間保存することも可能である。その最大にして最良の事例がウィキペディアであると言える。

2.5 メインカルチャーとサブカルチャー

ウィキペディア日本語版は、以前から、サブカルチャー関連の項目にあまりに偏りすぎているという指摘があった[木村08]。英語版とフランス語版と日本語版のウィキペディアを比較して、「こんなに違うウィキペディア」とのキャプション入りで、日本語版が（サブカルチャー系の編集が多くて）特殊だ、と主張している。これを受けて、匿名掲示板やニュースサイトでも、「日本のWikipedia 編集回数ランキング、アニメ・漫画が上位独占」「英語版とフランス語版では、政治的・宗教的著名人の項目や政治的な事柄の項目が上位にランクイン（ジョージ・ブッシュ／2006年イスラエル・レバノン紛争等）したのに対し、日本語版では上位5位が全て漫画・アニメに関するもの（ONE PIECEの登場人物一覧／舞-乙HiME等）となり、上位の項目はまるで異なる」などと話題になった（一部の顰蹙を買った）。

サブカルチャー関連項目の多さについては、筆者も今回じっくりと調べてみたが、たしかに事実である、と言える。たとえば英語版では、最も多く編集されている項目は「George W. Bush」であり、4万回以上編集されている。また、「Barack Obama」という項目はまだ18000回程度しか編集されていないが、Barack Obamaに関連する項目（たとえば、「Barack Obama presidential primary campaign, 2008（2008年バラク・オバマの主要な選挙活動）」、「Inauguration of Barack Obama（バラク・オ

バマの大統領就任式）」、「Barack Obama citizenship conspiracy theories（バラク・オバマの市民権陰謀論）」、などを総計するとGeorge W. Bushの最多記録にそろそろ及ぼうとしている。Bushの時代は皆が一つの項目に記事を書き込もうとしたのだろうが、Obama時代では一つの記事があまりにも肥大化しないように、いろいろな記事に分散して書くようになったのだろう。必ずしも編集回数を増やしすぎないようにする目的ではあるまい。ともかく、このように、英語版では政治や経済などのメインカルチャーに属する話題が非常に多い。

一方で日本語版では、最も編集回数が多いのは「ONE PIECEの登場人物一覧」であり、通算6000回程度である。2位以下は「仮面ライダー電王」「仮面ライダーディケイド」「舞-乙HiME」「銀魂」「BLEACH」「AV女優一覧」「銀魂の登場人物一覧」などが続く。少なくとも、編集回数という尺度で見ただけの場合には、日本語版の上位にくるのはテレビ、アニメ、漫画などのいわゆるサブカルチャーに属する項目である。

最近日本語版で、急速に編集回数を増やしてきた「仮面ライダーディケイド」という項目に着目して、なぜ日本語版ではこのようなことが起こるか、説明してみよう。初期の仮面ライダーをテレビで見たことがある、という程度の筆者らの年代の者にはわかりにくいのだが、仮面ライダーシリーズは、未だに毎年のように新作が放送される人気番組である。その主人公にはイケメン俳優が起用され、子供とその母親が主要な視聴層であると言う。

その視聴者が毎週新しい回が放送され、毎週新しい事実が加わるたびになんらかの書き込みをする。このために半年程度で数千件の書き込みに達する。しかしこの種のテレビ番組は半年か長くとも一年で放送が終了するので、それ以後はあまり編集回数が伸びない。長寿番組でも、ある一定の編集がなされると、それ以上はあまり伸びない、ということが日本語版では良く見られ、そのために数千回程度の規模のテレビ放送やアニメに関する記事が大量に存在することになる。

では、「仮面ライダーディケイド」を編集したユーザーの実数はどのくらいなのか。編集には、いわゆるIPユーザーと呼ばれる匿名ユーザーも加わるために確かなことは言えないが、登録ユーザー数は確定しており、135である。従っておそらくは、編集に関与したユーザー数はだいたい200人から400人くらいではないかと推測される。その他の日本語版の最上位にくる記事も同様なのではないか。つまり、ウィキペディア日本語版の目立つ傾向というのはたかだか数百人のコアなファンによって形成されているのである。日本人全体の人口に比べればあまりにも一部の人間たちだ。いわゆる組織票によって簡単にどうでも

なるレベルと言える。

これに対して「日本/log20090827」（2009年8月27日に「日本」からログとして分離された履歴）という項目は登録ユーザだけで594人が関わっている。編集回数は3600回とやや少なめだが、関わったユーザ数ではおそらく最多であろうと思われる（とは言え、たかが千人程度で1位になれるわけだ）。このように、計数する尺度によっても見えてくる傾向は変わって来て、歴史や政治などのいわゆるメインカルチャー系の記事が上位に来る。ちなみに編集に関わった登録ユーザ数の2位以降の順位は「日本の漫画家一覧（398人）」「大韓民国（395人）」「フジテレビジョン（391人）」「麻生太郎（383人）」「2ちゃんねる（362人）」「織田信長（358人）」「トヨタ自動車（355人）」「涼宮ハルヒシリーズ/log20090701（354人）」「第二次世界大戦（350人）」などとなっている。これらの人数は先ほど例に挙げた「仮面ライダーディケイド」の倍以上であって、いかに「仮面ライダーディケイド」などのサブカルチャー系の記事が少数の人たちによって（多くの場合は短期間に）頻繁に（おそらくは細部が）編集されているかがわかる。

他の言語版でも上位に芸能やサブカルチャーの項目が来ないわけではない。英語版では、編集回数2位は「プロレスラーの一覧（List of World Wrestling Entertainment employees）」だし、5位は「Michael Jackson」、7位は「Britney Spears」、8位は「Wii」、12位は「Rune Scape（オンラインのRPG）」である。プロレスラー一覧は3万回以上、マイケル・ジャクソンやプリトニーなどは2万回編集されており、日本のサブカルチャーよりも桁違いに多いので、簡単な比較はできない。ちなみに英語版で過去半年以内で編集に関与した登録ユーザ数の順で上位の記事は次のようになる。

- Michael Jackson (1101 users) マイケル・ジャクソン。
- Deaths in 2009 (913 users) 2009年の死亡者一覧。
- Air France Flight 447 (636 users) 2009年1月に起きた飛行機事故。
- 2009 flu pandemic (519 users) 2009年インフルエンザ流行。
- Barack Obama (487 users) 現職大統領。
- Death of Michael Jackson (389 users) マイケル・ジャクソンの死
- 2009 Iranian election protests (364 users) 2009年イラン選挙後の抗議活動。
- Inglourious Basterds (362 users) 新作映画。
- Harry Potter and the Half-Blood Prince (film) (358 users) 新作映画。

- District 9 (354 users) 新作映画。

英語版でも短期間で見れば、歌手や映画などの話題が上位に来やすい。しかし、長期的には、いわゆるメインカルチャーと言われる政治、宗教などの話題が総和として大きくなる。日本語版は逆に、長期に渡ってじわじわ編集回数が増えていく話題が少ないのかも知れない。

ドイツ語版で編集回数の2位は「ハリーポッターの登場人物一覧」で6000回程度、3位は「シンプソンの登場人物一覧」であって、上位にサブカルチャー系の項目が混じっているという意味では日本語版と同じである。ちなみに1位は「ドイツ」、4位は「ナザレのイエス」、5位は「エホバの証人」である。フランス語版でも3位はプロレス、6位はテニス選手、7位はマイケル・ジャクソン、8位はビートルズなどとサブカル系が上位に混じっている。ちなみに1位は「2005年2月」、2位は「ニコラ・サルコジ」、4位は「アルジェリア」、5位は「オラン（アルジェリアの都市）」である。

日本語版の特徴として挙げてよいと思われるのは、編集回数100回から2000回レベルに大量に作られた鉄道関係の項目群であろう。旧国鉄の電車や現在の私鉄の電車に至るまで、膨大な数の項目がある。駅についてもまた同様である。地方都市だと、駅に関する項目があっても地方自治体についての記事がないことも多い。個々の項目の編集回数は少なくても、関連項目全体ではどれほど編集されているかわからない。このような傾向はおそらく他言語版には見られず、いかに日本人が鉄道が好きか、日本語版が「オタクの百科事典」と化しているかの証拠となるだろう。

ごく一部の例を挙げると、「国鉄485系電車」は2317回編集されており、編集に関わった登録ユーザは130人である。また、「新幹線 N700系電車」は905回編集されており、関わった登録ユーザは139人である。この事例を見ても「国鉄485系電車」は編集回数は多いがユーザは少なく、「新幹線 N700系電車」は編集回数は少ないがユーザがやや多いのであって、どちらがより人気が高いのか、簡単には決められない。

このように個々の記事の編集回数は少ないが、類似記事が膨大にあるという現象、つまり「趣味の細分化」現象が、日本語版のオタク系の話題に顕著に見られる、ということは確かだろう。こまごまとしたたくさんの項目を作りたがるので、突出した記事ができにくい。また、政治や経済の記事をこまごまと執筆しない、あるいは編集が集中しにくい、ということはある。しかし現時点でこれは筆者の大まかな印象に過ぎず、正確な全貌はさらなる解析を行う必要がある、と言える。

3. ウィキペディアのデータ解析

ウィキペディアのデータを解析するには、およそ三つの方法がある。一つは google などの検索エンジンがやるように、HTML を蒐集して解析するというものであるが、これはシステムに負荷をかけるという意味で、ウィキペディアではまったく推奨されておらず、本稿でもこれ以上言及しない。また一つはすでに述べてきたように、無償公開されているダンプデータを解析するというものである。もう一つのやり方としては、最近更新された箇所が RSS で配信されるので、これを解析するというものである。

3.1 RSS の処理

まず RSS を解析する手法について述べる。通常ダンプデータは 1 週間から 2 週間に 1 回程度の頻度で作成される。一方 RSS は、最近数分間の編集履歴などが得られるので、どちらかと言えばどこかで編集合戦が起きているとか、あるいは今日一日人気だった項目などを抽出するには適しているだろう。このようなサービスを提供しているものとしては wikirage[[Wikir](#)]や wikichecker[[Wikic](#)]などがある。これらは 5 分に一度程度の頻度で Wikipedia から更新情報の RSS を取得し、それを解析した結果を随時 Web 公開する。

管理者にとっては、削除依頼や保護依頼、今まさに頻繁に編集されているページ、あるいは新規に作成されたページなどをできるだけすばやく RSS で知る必要がある。あるいは、スパムに類する書き込みが行われたかどうかもある程度 RSS でチェックできる。また、一般のユーザも、自分が執筆に関わったページが書き換えられたことを RSS で直ちに知りたいと思うだろう。

だが、このような RSS 配信に基づく解析結果は、1 日とか 3 日程度の期間の編集合戦が検出されるのみであり、1 週間よりも長い期間にわたる、数十件ほどまとまった編集の変動や動向については、やはりダンプデータを解析して観測した方が有用であると言える。

3.2 ウィキペディアダンプデータの処理

ウィキペディアが公開しているダンプデータ [Dump] にはいくつかの種類がある。完全な履歴を含んだデータとしては各言語に pages-meta-history.xml というものが用意されている。拡張子名からわかるように、このファイルは XML (extensible markup language) という形式で、つまりマークアップされたテキストファイルである。すべてを含んでいるとは言っても、登録ユーザの個人データや、特定版削除された版のデータは含まれていない。

日本語版の pages-meta-history.xml の全体のサイズは、未圧縮でほぼ 300 GB ある。これを最近の 7 zip というアーカイブを用いれば、わずか 2~3 GB に圧縮できる。この圧縮されたデータをダウンロードしてきて、解凍し、再びデータベースサーバに読み込めば、本家のサーバと同じようにデータをクエリ (取り出したり格納したりといった問い合わせ命令を送信) できるのであるが、このような大きなデータをオンライン化するには膨大な時間とマシンパワーが必要なので、通常はオフラインで解析を行う (ここで「オンライン」とはデータがデータベースサーバに格納されていて、クエリができるようになっている状態を言う。Web サービスとして公開されている状態と言ってもよい。一方オフラインとはデータベースサーバからダンプされたファイル形式のデータの状態を言う。本稿では以後そのような意味に用いる)。

また、データは XML で書かれているが、XML パーザ (XML を解析してくれるソフト) に直接 300 GB もの XML 文書を読み込むのは自殺行為に等しい。したがって、普通は一行ごとに読み込んで、必要なキーワードだけを抜き出してメモリに格納するという地味な作業を行う。実際、このダンプファイルは非常に巨大ではあるが、構造は至ってシンプルであり、階層は深くないので解析はさほど困難ではない (XML はタグを入れ子にしてデータを階層化できる。複雑に分岐した深い階層のデータは一般に解析に手間がかかる)。

英語版のウィキペディアには 2008 年 1 月現在で約 215 万項目の記事があり、日本語版のおよそ 5 倍の規模である。その履歴を含めたダンプファイル pages-meta-history.xml を 7 zip で圧縮したもので 17.3 GB あり、解凍すると 2.8 TB になるらしい。ちなみに George W. Bush の項目だけでも非圧縮 XML 形式で 3 GB に達する (つまりこの単一項目だけでも XML パーザによる解析は事実上不可能ということなる)。また、ファイルサイズの大きさもさることながら、ダンプデータをファイルに書き出し圧縮するだけであまりにも時間がかかってしまうために (驚くべきことに二ヶ月くらいかかるようだ)、現時点ではこのデータはアーカイブとして公開されておらず、入手可能な最近のものは上記のとおり 20080103 までである。

stub-meta-history.xml というファイルは、pages-meta-history.xml から本文だけを欠落させたものである。従って編集回数を知りたいとか、編集した時刻やユーザ名を知りたいなど、編集履歴だけを知りたい場合はこの stub-meta-history があれば十分である。またファイルサイズもほどよくコンパクトで、英語版を含むすべての言語版で入手可能である。そこで筆者は日本語版だけは

pages-meta-history を使い、英語版やドイツ語版などは stub-meta-history を使うようにしている。ちなみにウィキペディアで stub とは項目だけとりあえず立ててあるが中身のないページを本来は意味する。

4. 推薦システムの応用

4.1 概要

「推薦システム」(recommender system)「推薦エンジン」(recommendation engine)あるいは「協調フィルタリング」(collaborative filtering)などと言った手法は、商品(など)に対する他人と自分の評価の類似度を指標として、自分が未知の商品を購入するかどうかの判断の助けにする仕組みである。比較的少数のデータを対象とした研究がかなり古くから行われてきた形跡があるが、今日、この手法を有名にしたのは、オンライン書籍販売のアマゾン・コムである。アマゾンのサイトに行くと「この商品を買った人はこんな商品も買っています」と言った「お薦め機能」が導入されている。他の読者・購買者がどんな本を併せて買っているかを知ること、類似の本を次々にたどって調べることができる。著作者や出版社、あるいは書店などが恣意的に分類したのではなく、ユーザの購買履歴によって統計的に分類がなされるというところが、まったく新たな独立した尺度として役に立つ。

さてこの「推薦システム」だが、はてなアンテナの「おとなりアンテナ」或いは「おとなりページ」にも応用されている。はてなアンテナとは Web 日記に特化したリーダーの一種であるが、他人が公開しているアンテナと自分のアンテナに登録している項目が一致している度合いによって、自分のアンテナにはないが読んだ方がよい日記というものを、重複が多い順に推薦してくれる。ほかにも、「2ちゃんねる」のスレッド表示で「このスレを見ている人はこんなスレも見えています。」などと他のスレッドを紹介してくれるのも同じ仕組みの応用である。YouTube やニコニコ動画などの動画配信サービスの「関連動画」も同様の仕組みを利用して選ばれている可能性がある。

はてなには似たようなサービスとして「はてなブックマーク」通称「はてぶ」がある。自分が面白かった記事に公開ブックマークを付けるのだが、一種の人気投票システムと言え。アマゾン・コムにも書評という人気投票システムがあるが、「推薦システム」が「人気投票システム」と本質的に異なるところは、その「作為性」「恣意性」であると言える。人気投票とは本来恣意的なものであって、作為的に順位を操作しようとするものである。「推薦システム」は、前述したように、ユーザが

自発的に推薦するのではなく、「顧客情報の持ち主」がその顧客情報を解析した結果に基づいて推薦するのである。もちろん意図的に編集回数を増やして「人気投票」することもできなくはない。意図的に多くのユーザ登録をして編集に関わったユーザ数を水増しすることもできなくはないだろう。このような恣意性を完全に排除することはできないが、通常、顧客が蒐集されつつある自らの情報全体を操作することは困難であって、故に推薦システムの結果はユーザの作為のない行動履歴から得られるとみなすことができる(推薦者が商売の都合で恣意的な推薦を混入させる可能性もあるが、ここでその問題は無視する)。

「推薦システム」は「商品」と「購買者」の関係が一番わかりやすいが、それ以外の関係にも応用が利くことがわかる。抽象化してみると、「トピック」の集合と「ユーザ」の集合があって、原則として「トピック」は Web に公開されているが、「ユーザ」は個人情報として秘匿されているという状況があるとする。ユーザはトピックを好みによって取捨選択することができる。トピックは商品であったり、単なるブックマークのようなものであったり、あるいは閲覧履歴であったりする。このようなトピックとユーザの関係があった場合に、トピックどうしの類似関係が、トピックを共有するユーザの重複度から得られるということになる。

わかりやすいように、例を挙げて説明しよう。図1において、ユーザ1はトピック2とトピックmを編集しており、ユーザ2はトピック1を編集し、ユーザ3はトピック4を編集し、ユーザ4はトピック2と3とmを編集し、ユーザnはトピック1と3とmを編集してい

		トピックID							
		1	2	3	4	...	m	...	
ユーザID	1		●					●	
	2	○							
	3				○				
	4		●	●				●	
	...								
	n	●		●				●	
	...								
	...								

図1 トピックとユーザのマトリクス

るとする。ここでトピック m に着目すると、トピック m を編集したユーザは 1 と 4 と n である。ゆえに、トピック m を編集したユーザのうち、トピック 2 またはトピック 3 も編集したユーザは二人ずつ居て、トピック 1 を編集したユーザは一人居ることがわかる。同図中、白丸はトピック m の推薦に無関係な編集を、黒丸はトピック m に関係のある編集を表す。

このように「推薦システム」は「トピック」と「ユーザ」のマトリクスから得られるが、「トピック」の数が増えれば当然ながら「ユーザ」の数も増えて、しかもこのマトリクスは極めて疎になるから、実際にマトリクスをメモリ上に構築するのは効率的ではない。そこで、メモリ上には「トピック」の集合と「ユーザ」の集合を保持し、各トピックにはそのトピックに関与したユーザへのポインタを持たせ、また各ユーザにはその関与したトピックへのポインタを持たせる。

そうすると、一つのトピックに関与するすべてのユーザが得られ、それらのユーザが関与したすべてのトピックが重複を含めて得られる。従って一つのトピックに関与する「推薦トピック」群が得られる。ここで一つ重要なことは推薦トピックが得られた後にはユーザ情報をすべて破棄してよく、参照する必要もないということである。

この一連の処理は極めて時間がかかる。たとえばトピック群とユーザ群だけをデータベースにオンラインで持っていたとして、web サーバにリクエストしたときに「推薦トピック」を検索したとすると、結果が出るまで何分も待たされることになり、実用性に乏しい。時間がかからないように検索対象を絞り込むと結果は著しく不正確になる。そこで、オンライン化するには、各トピックに対してあらかじめオフラインで推薦トピック群を計算しておく必要がある。このような前処理は Web 上で実時間で処理結果を表示するアマゾン・コムやはてな、2ちゃんねるなどの推薦システムでも当然採用されているものと思われる。

4.2 ウィキペディアデータベースに対する推薦システムの導入

Wikichecker[[Wikc](#)]はウィキペディアに「推薦システム」をすでに導入しているサービスの一つであるが、オフラインの前処理を含まないために、非常に低速であるし、またその結果は編集回数が多い上位ユーザに関与するトピックに限られている。またデータベースサーバにかなり負荷をかけるために、アクセスが多いサービスには向かない。筆者は4.1で述べたオフラインの前処理を導入することによって、オンライン化に十分な処理速度

を実現した。

ウィキペディア英語版のように記事数が多いところはユーザ数も多く、記事とユーザのマトリクスも巨大になる。ユーザ数が記事数に比例すると仮定すると、このマトリクスのサイズは記事数の自乗に比例すると言える。そのため、現実的な処理時間を実現するには記事の絞り込みが有効であり、推薦記事の計算には、現在、最近六ヶ月に編集した登録ユーザだけをカウントするようにしている。これによって現時点で記事数が少ない韓国語版や中国語版などの言語版ではお薦め記事の計算は Dell PowerEdge R200 上で数十分で終了するが、英語版は約 1 日かかる。編集時期の制限を設けない場合、日本語版で約 3 日かかる。また、編集回数が 10 回未満のページは項目だけの形式的な記事（スタブ、または単なるリダイレクト）である場合がほとんどであるので除外する。

4.3 人気記事を別テーブルにすることで Web 表示を高速化

データベースの応答速度はテーブルのサイズに大きく依存する。1000 件の項目しかないテーブルから結果を検索する場合と比べて、100 万件の項目を含むテーブルから検索する場合は当然処理がかなり遅くなる。

そこで筆者は、編集回数、および編集に関わった登録ユーザ数の上位の記事を抜き出して「人気記事」として別テーブルを用意した。まずこの「人気記事」テーブルから先に検索し、このテーブルに無いときだけ全ての記事が格納されたテーブルを検索する。この手法は体感速度を劇的に改善する。

本稿の最後に現在実際に公開している推薦システムの Web ページのようすを載せる (図 3)。

4.4 差し戻し回数のカウント

差し戻し (リバート、revert) とは、主にウィキペディアで使われる、リビジョン管理に関する用語であり、スパム、荒し行為などのある版を履歴に残したまま、過去の妥当な版をコピーして最新版に復帰させることである (スパムや荒らし程度であれば、著作権や個人情報の問題に抵触しないので、特定版削除はされない。その代わり、過去の履歴に膨大な罵詈雑言が残されていることがある)。リバートが多い項目には、単に荒しやスパム行為が多い、あるいは IP ユーザなどの初心者による編集が多い、見解が一致しにくい、編集合戦が起きやすいなどといった原因があると考えられる。Wikichecker[[Wikc](#)]では編集合戦センサーに最近のリバート回数を用いている。

筆者はリバート回数の総数をカウントしてみた。英語

版の結果：

- George W. Bush (581回、1.3%)
- List of World Wrestling Entertainment employees (313回、1%)
- Windows Vista (257回、2.2%)
- Grunge (245回、6.1%。音楽ジャンルの一つ)
- Wikipedia (242回、0.8%)
- Adolf Hitler (231回、1.2%)
- Jesus (216回、1%)
- World War II (204回、1%)
- Wii (193回、1%)
- Anarchism (192回、1.3%)

日本語版の結果:

- 西日本旅客鉄道 (32回、2.1%)
- 破面 (25回、0.8%)
- 日本/log20090827 (25回、0.7%)
- 鉄道事故 (18回、0.9%)
- モーニング娘。 (18回、0.6%)
- 山手線 (18回、1.5%)
- ふたりはプリキュア (17回、1.2%)
- 東京急行電鉄 (15回、1.4%)
- ウィキペディア (15回、2%)
- 大韓民国 (15回、0.7%)

リポートが多い記事の順位は、明らかに他の指標の順位とは異なっており、独自傾向が認められる。しかし、リポート率はだいたい1%未満であって、リポートという行為がただちに加熱した編集合戦に結びつくわけではないことがわかる。

たとえば、前述の「仮面ライダーディケイド」だが、なんとリポート回数は0である(ただし著作権侵害による特定板削除は何度か行われている)。日本語版ではリポートのような編集合戦がまったく起きないままに、数千回の編集が粛々と行われているらしいことが推測できる。

4.5 ガジェットやブログパーツとしての提供

Web 公開するにあたって、図2のように、google gadget やブログパーツとしてもコンテンツを提供している。ウィキペディアの「おまかせ表示」に「推薦システム」を導入したガジェットと言え。メニューで言語を選択することもできる。

4.6 他手法との比較

ウィキペディア以外の組織でウィキペディアのデータ

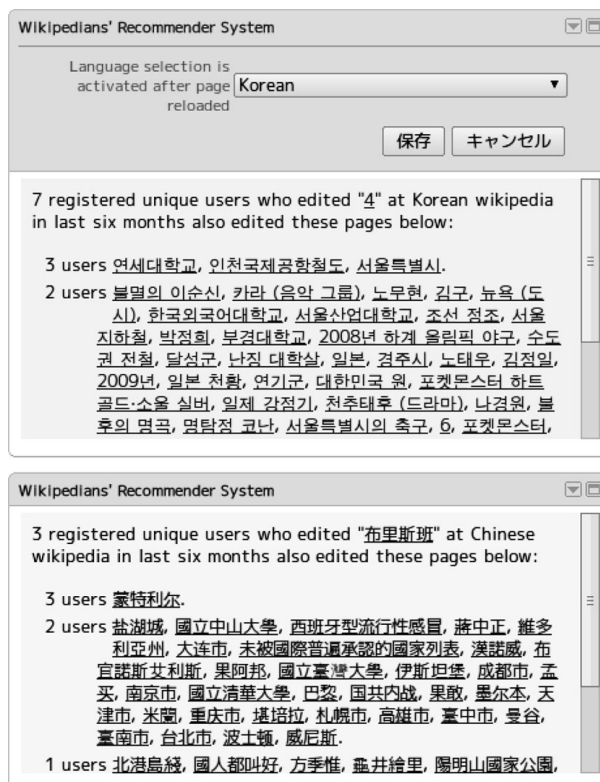


図2 google gadget の例

ベースを解析する Web サービスとしては、Wikiscanner [Wiks] や Wikirage [Wikr]、Wikichacker [Wikc] などのように、ある特徴的な振る舞いをするユーザを見つけ出して、その IP アドレスから可能であれば所属をつきとめたり、或いは編集履歴を暴露するようなものが多い。これらは、特定の組織、特に政府機関やマスコミなどが恣意的な編集を行った履歴を明らかにするためであったり、荒らし行為や編集合戦などを検出して保護や半保護などの対策を講じるためであったりする。また、不良ユーザを特定してブロックすることにも利用される。実際、Wikiscanner によってマイクロソフトとソニーの間の泥試合がやり玉にあがったし、宮内庁が天孫陵について、厚生労働省が特定代議士について編集を行ったことなどが明らかにされた[山本・古田08]。他にもちょっとネットで検索するだけでさまざまな疑惑が見つかるだろう。これらの機能はもちろんウィキペディアにとって有益であり必要であるが、「推薦システム」として考えた場合には不要であるし、むしろ有害でさえある。センセーションリズムと実用性とは必ずしも一致しない。

「推薦システム」は通常は商業サイトなどで使われるものであり、顧客情報は最高機密である。ユーザ情報を完全に隠蔽しつつ、どの商品がどのくらいどういうタイプのユーザに好まれているかを提示するのが「推薦シス

テム」の最大の特徴である。そうしなければユーザからのクレームを避けることはできないし、ユーザは安心してサービスを利用できない。

また、ウィキペディアのデータベース検索を他のサービスと抱き合わせで提供するものや API として提供するものもある[Goo][Sim]。これらはウィキペディアの記事本文を主要なサービス対象とするが、同時にウィキペディアが潜在的に抱えている著作権侵害や個人情報保護の問題を自らも抱え込むことになる。「推薦システム」はウィキペディアの本文とは本質的に無関係であり、いわばメタ情報だけから抽出される情報であり、それらの問題に関係する確率はかなり低減されていると言ってよいと思う。

ウィキペディアの管理者らの関心事は主として、信頼に足る、正確な内容のオンラインの百科事典を作ること、である。しかし、筆者が提案する手法は「本文の信頼性」という属性にはまったく関知しない。百科事典としての質の向上という命題は、ウィキペディア自身に取り組んでもらうしかない。また、「刹那的な編集合戦」についても本手法は関知しない。各言語版におけるもっと長い期間に渡る編集傾向、および言語版の間の比較、といったことに興味がある。そこから浮かび上がってくるのは、日本語版の特殊性であるとか、韓国語版や中国語版との関連であるとか、それら諸地域がどのようなことに興味を持っているのか、などである。

4.7 推薦の傾向

提案手法を理解してもらうには筆者らが公開しているサイト[Wpw]を訪れてもらい、自由気ままにブラウズしてもらうのが一番良い体験方法であるが、その中の特徴的な結果を以下に列挙しておく。

ウィキペディア日本語版で「日本/log20090827」を編集した登録ユーザは594人であり、それらのユーザ編集した他のページは以下の通り：

- 111人 アメリカ合衆国。
- 104人 大韓民国。
- 96人 中華人民共和国。
- 85人 第二次世界大戦、天皇、イギリス。
- 83人 東京都。
- 82人 小泉純一郎。
- 81人 日本語。
- 80人 昭和天皇。
- 79人 ドイツ。
- 78人 朝鮮民主主義人民共和国、フランス。
- 77人 東京。
- 76人 麻生太郎。
- 73人 織田信長。
- 72人 太平洋戦争、ドラえもん。
- 70人 イタリア、インド。

同様に「アメリカ合衆国」を編集した登録ユーザは310人で、次のようになる：

- 111人 日本/log20090827。
- 74人 第二次世界大戦。
- 73人 フランス。
- 72人 中華人民共和国。
- 69人 ドイツ。
- 67人 イギリス。
- 63人 大韓民国、イタリア。
- 61人 アメリカ同時多発テロ事件。
- 60人 オーストラリア。
- 56人 カナダ、ジョージ・W・ブッシュ。
- 54人 織田信長、インド、昭和天皇、小泉純一郎、オーストリア。
- 53人 天皇。
- 52人 東京都。

このように、日本語版で「日本」と「アメリカ合衆国」は相互に編集している共通ユーザが多数存在し、またその他に編集しているページもメインカルチャーに属する傾向がみられる。いわばメインカルチャーに属する項目が「芋づる式」に抽出されていると言える。

次に、典型的なサブカルチャーの項目を見てみよう。「ONE PIECEの登場人物一覧/log」を編集した登録ユーザは112人で、他にも編集したページは以下の通りである：

- 55人 ONE PIECE/log。
- 51人 海賊 (ONE PIECE)。
- 36人 海軍 (ONE PIECE)。
- 35人 モンキー・D・ルフィ。
- 31人 ONE PIECEの用語一覧、バロック・ワークス。
- 29人 NARUTO-ナルト-の登場人物一覧。
- 28人 悪魔の実。
- 27人 DEATH NOTEの登場人物。
- 26人 ウソップ。
- 25人 空島 (ONE PIECE)、サンジ。

このように、NARUTO や DEATH NOTE など比較的傾向が近い他のアニメ・漫画作品も混ざってはいるが、ほとんど完全に ONE PIECE に関連する項目が抽出されている。次に「仮面ライダーディケイド」を編集した登録ユーザは135人で、他にも編集したページは以下の通りである：

- 45人 仮面ライダーキバ。
- 40人 仮面ライダーカブト。
- 37人 仮面ライダー龍騎。
- 35人 仮面ライダー電王、侍戦隊シンケンジャー。
- 33人 劇場版 仮面ライダーディケイド オールライダー対大ショッカー、仮面ライダー W。
- 31人 仮面ライダー剣、炎神戦隊ゴーオンジャー。
- 29人 獣拳戦隊ゲキレンジャー。
- 28人 仮面ライダーアギト、仮面ライダークウガ。
- 27人 仮面ライダー電王/log20080331。
- 26人 仮面ライダー響鬼、仮面ライダーシリーズ、スーパー戦隊シリーズ。
- 25人 仮面ライダー。
- 24人 仮面ライダー555、機動戦士ガンダム00、劇場版 仮面ライダー電王 俺、誕生!。
- 23人 ウルトラマンメビウス。
- 22人 新世紀エヴァンゲリオン、仮面ライダー BLACK RX。
- 20人 仮面ライダー X。

この場合もやはりウルトラマンやエヴァンゲリオンなどもまざってはいるが、ほとんどすべて仮面ライダーシリーズに関連する項目が抽出されている。比較のために中国語版の「仮面ライダーディケイド」の結果を示す：

「幪面超人 DECADE」 28人

- 8人 幪面超人龍騎（仮面ライダー龍騎）。
- 7人 假面騎士 Double（仮面ライダー W）。
- 6人 侍戦隊真劍者（侍戦隊シンケンジャー）。
- 5人 假面騎士系列（仮面ライダーシリーズ）、幪面超人 KABUTO（仮面ライダーカブト）。
- 4人 面超人電王（仮面ライダー電王）、幪面超人古迦（仮面ライダークウガ）、幪面超人 KIVA（仮面ライダーキバ）、動漫角色生日列表（アニメキャラクターの誕生日一覧）。
- 3人 傲嬌（ツンデレ）、守護甜心！（しゅごキャラ!）、守護甜心！角色列表（しゅごキャラ！キャラクター一覧）、絶対武力 Online（カウンターストライク Online）、幪面超人響鬼（仮面ライダー

響鬼）、機動戦士 GUNDAM 00登場機體列表（機動戦士 GUNDAM 00に登場する機体一覧）、7-Eleven（セブンイレブン）。

前述のように中国語版（実質的には台湾・香港版）や韓国語版は項目数・編集回数ともに現時点では非常に少なく、さらにサブカルチャー系の項目には日本語版の記事を翻訳しただけに等しいものが相当見られる。編集回数が不足していると共通ユーザ数も少なくなり、その結果ユーザ数が少ない下位の抽出結果にはノイズが多く含まれるようになる。ただし単なるノイズなのではなく、仮面ライダーシリーズそのものではないが、「ツンデレ」「しゅごキャラ!」「カウンターストライク Online」「機動戦士 GUNDAM 00」といったように、比較的近接するサブカルチャーの領域に属する項目が抽出されていることがわかる。かなり少数の編集者の趣味を反映したものである可能性も高い。いずれにしても、上記の結果を見ただけでも台湾や香港におけるサブカルチャーのおよその傾向は見て取れるだろう。

英語版でマイケル・ジャクソン、バラク・オバマ、ジョージ W ブッシュ、プロレスラー一覧について得られた結果を次に示す：

「Michael Jackson」 1101人

- 124人 Death of Michael Jackson.
- 78人 Deaths in 2009.
- 65人 Barack Obama.
- 55人 Michael Jackson memorial service.
- 54人 Farrah Fawcett.
- 47人 Madonna (entertainer).
- 46人 Britney Spears, Lady Gaga.
- 45人 Sonia Sotomayor, The Jackson 5.
- 44人 Air France Flight 447, Paul McCartney, Billy Mays.
- 43人 2009 flu pandemic.
- 42人 Ted Kennedy, Thriller (album).
- 41人 2009, Patrick Swayze.

「Barack Obama」 487人

- 65人 Michael Jackson.
- 55人 Ted Kennedy.
- 44人 Barack Obama citizenship conspiracy theories, Presidency of Barack Obama.
- 43人 George W. Bush.
- 42人 Sonia Sotomayor.
- 38人 Deaths in 2009.

- 37人 2009 Nobel Peace Prize.
- 35人 Sarah Palin.
- 33人 2009 flu pandemic.

「George W. Bush」168人

- 43人 Barack Obama.
- 24人 Bill Clinton.
- 23人 Ted Kennedy.
- 21人 Michael Jackson.
- 19人 2009 flu pandemic, Sonia Sotomayor.
- 17人 Mahmoud Ahmadinejad, Abraham Lincoln.
- 16人 Deaths in 2009, John F. Kennedy, George H. W. Bush, Jimmy Carter, Franklin D. Roosevelt, Sarah Palin, Ronald Reagan.

「List of World Wrestling Entertainment employees」139人

- 33人 CM Punk (プロレスラー).
- 28人 Adam Copeland (プロレスラー), List of Total Nonstop Action Wrestling employees.
- 26人 Paul Wight (プロレスラー).
- 25人 The Legacy (professional wrestling), Chris Jericho (プロレスラー).
- 23人 Night of Champions (2009), Dave Batista (プロレスラー), Jason Reso (プロレスラー).
- 22人 John Cena (プロレスラー).
- 21人 The Undertaker (プロレスラー), SummerSlam (2009), World Wrestling Entertainment, Curt Hawkins and Zack Ryder (プロレスのタッグチーム).

ここで筆者は上記結果について、想像力を駆使して大胆な仮説を立てていこうと思う。あくまでも推測であって、筆者が引いた補助線が正しいかどうかは必ずしも保証はできないので、判断の一材料としていただきたい。まず、Michael Jackson の結果において、Farrah Fawcett、Madonna、Britney Spears、Lady Gaga は女優または女性歌手であり、Sonia Sotomayor は最高裁判所の陪席判事、Billy Mays はテレビ通販のセールスマン、Patrick Swayze は男優である。思うに、Michael Jackson と Farrah Fawcett、あるいは Billy Mays、Sonia Sotomayor との論理的な関連性は極めて低く、百科事典の説明で一方が他に言及されたり、同一カテゴリーに分類される可能性はほとんどない。また日本人は Michael Jackson は知っていても、残りの三人についてはほとんど誰も知らない。しかし、アメリカ人にとってこの四人は、テレビ

に良く露出する著名人として同じ範疇に属しているのである。少なくとも、我々は Michael Jackson の編集を好む人たちは、テレビで良くみかける他の有名人の編集も（おそらくはかなり安直な、ゴシップ的な理由で）好むであろうということを、一つの客観的な事実として知ることができる。

推薦システムが、単に有名人から順に抽出されるランキングのようなものであるかと言えば、そうとは言えない。たとえば、2.4節に示したような、「Michael Jackson」「Deaths in 2009」「Air France Flight 447」「2009 flu pandemic」「Barack Obama」といったような抽出結果が典型的なランキングの結果であり、単にある指標値に基づいて、人気がある順にほとんど無関係にトピックがピックアップされている。推薦システムの結果にはもっと強い「何か」の関連性が見られるのが普通である。

アマゾン・COMの推薦システムでも同様に観察される現象であるが、たとえば「広辞苑」を買ったユーザは「現代用語の基礎知識」など「辞典」という同一カテゴリーの書籍も買う傾向があるが、同時に「フェアブル昆虫記」のような別カテゴリーの書籍も買っている。「辞典」というのは図書館学的な分類であるが、同時に「児童図書」あるいは「教育図書」という別の分類もあり得て、これはこれで一つのリーズナブルな分類である（少なくとも書籍のマーケティング的には）。

このように推薦システムはある与えられたカテゴリーに基づいた分類をするのでもなければ、ある尺度によってランキングを出してくれるのでもない。いわば、多くの人の個人的嗜好の総和に基づく、あらかじめ明確に与えられてはいないいくつかのカテゴリー分けが複合した結果を提示してくれるのであり、それによって今まで気づかなかった分類方法を気づかせてくれることもあり得るのである。それら複数のカテゴリーは狭いカテゴリーから広いカテゴリーへと包含関係になっているかもしれないし、異なるカテゴリーの共有部分かもしれない。

Barack Obama の結果は、有名人または有名な事件であるという点以外にあまり共通性が見いだせない。これはつまり Barack Obama という人における最も重大な属性が「アメリカ的に重要」だということだからかもしれない。つまり、Barack Obama を編集したい人は、何か国家的に大きな話を編集したい人たちのだろう。

George W. Bush の結果からは、歴代の有名な政治家がピックアップされているのを見て取ることができる。

List of World Wrestling Entertainment employees の結果ではみごとにまでにプロレスラーばかりが抽出されている。またその中でもどのレスラーに人気があるのかということが、この結果からおぼろげにわかる。ウィキペ

ディア本文は百科事典の性格上、そうした編集者の主観をそのまま記述するものではなく、ある程度整理された結論のみが叙述され、執筆意図は隠されている。プロレスラーの一覧は一覧であって人気ランキングではないが、編集者の執筆動機にはやはりひいきがあるのであって、その傾向を間接的に知ることができると言える。

5. おわりに

本稿執筆段階では、まだこのシステムは公開して数週間しか経っておらず、有用であるかどうかの具体的なデータは得られていない。本システムは、Webに公開して、多くのユーザに利用してもらって初めてその有用性が立証できる。ウィキペディアンらの編集履歴などを公開することなく、またウィキペディア本文の著作権問題に抵触することなく、ウィキペディアをより便利に使えるようにするというのが本システムの目標である。本サイト[Wpw]にグーグルなどの検索エンジンからのアクセスがようやく増えつつあるが、この先どの程度利用されるかについては別の機会に述べてみたい。

本稿は思いの外長文になってしまった。その総論も煩瑣になってしまうために、以下、個別事項ごとにまとめを述べる。

5.1 サブカルチャーについて

筆者は、ウィキペディアのデータベースを解析していて、メインカルチャーとサブカルチャーの違い、サブカルチャーがメインカルチャーに変化していく過程、などについて数理的に説明できるのではないか、と思った。数千回の編集記録を残し、その間ほとんどリポートなどの編集合戦を経験せず、IPユーザが多く、かつユーザ総数が比較的少ない場合には「自作自演」が疑われるが、その可能性を含めて、統計的にこのような現れ方をするものがサブカルチャーの典型であると仮定しよう。日本語版ではこのようなものはアニメやバラエティ番組などに数多く見られる。次に、英語版の中でこのような統計的特徴を持った記事を探してみると、非常に多く見られるのだが、たとえば「List of character from Total Drama Island」などがその典型的な例と言える。Total Dramaとはアニメ番組であり、Total Drama Islandはその中の1シーズンに当たる。つまりTotal Dramaはいくつにも分けて放送されたシリーズものであり、Total Drama Islandは半年程度で放送され、完結した番組である。編集回数は10982回に及ぶがIPユーザ率が77.8%に達する。

思うに、信号理論風に言えば、サブカルチャーとはインパルス応答であり、過渡現象なのであって、故に一過性のもので高周波成分が多い。これに対してメインカル

チャーは定常状態なので必然的に低周波成分が多く、かつ長期的なのである。サブカルチャーはやはり廃りの流行のようなものと言って良いかも知れない。少人数でも一時的な流行は作り出すことができるが、パワーの総和が少ないので、長期的に持続できない。

英語は話者人口が多いので、低周波成分が細かなノイズ(サブカルチャー)を覆っている。しかし、サブカルチャーが存在していないわけではなく、単に目立たないだけなのである。言語や宗教などは数億人から数十億人の人口を擁しているから、国家よりも文化に与える影響が大きい場合があり得る。英語版ではこの影響が圧倒的なために、メインカルチャーが主流であるように見える。一方で日本は、宗教も言語も国家も、それほどまで大きくはない。そもそも、メインカルチャーに属する人たちがウィキペディアの編集に関わってないのだろう。そうすると、高周波成分であるサブカルチャーが目立ってくる。しかし、単なるサブカルチャーで終わる項目は、編集回数が数千回程度で止まってしまう、ユーザの厚みがない(ユーザ数が数百人しかいない)ので、積み増しすることができない。他方でメインカルチャーに属する項目は、毎年毎年少しずつ成長して行くので、やがて日本語版でもサブカルチャー的な項目群を覆い隠すほどになるに違いない。

江藤淳はサブカルチャーをトータルカルチャー(全体文化)に対する部分的な文化として言及しているが[大塚04]、ウィキペディアの解析結果においても、サブカルチャーをそのような意味でとらえた方が理解しやすいと言える。

ある社会において支配的な文化、つまりメインカルチャーに対して、娯楽的なもの、同人的なもの、反体制的なもの、アンダーグラウンド的なもの、草の根的なもの、珍奇で目新しいもの、若者文化的なものなどをサブカルチャーと呼ぶ傾向があるが、支配的な順でいえば、米国のサブカルチャーの方が日本のサブカルチャーよりも圧倒的に巨大だし、日本のメインカルチャーは(少なくともウィキペディア上の見かけでは)それらよりも小さい存在であって、決して支配的とは言えないのである。また日本のサブカルチャーの中には世界的に支配的なものも少なくない。またある種の伝統芸能や古典文学はメインカルチャーと見なされているが、ほとんどの日本人にはまったく関心がない。これらの現象をうまく説明するにはどうすればよいか。

ここで、江藤淳のように(その規模の大小はともかくとして)、「ある民族や言語の話者全体が共有する文化」をトータルカルチャーと呼び、それに対して一部の「隔離されたコミュニティの中だけで通用する文化」をサブ

カルチャーと位置づけたとする。進化論において、突然変異は隔離された狭いコミュニティの中でより頻繁に発生する。サブカルチャー集団の中で生まれた新しい文化はトータルカルチャーの中に取り込まれることによって淘汰され洗練されて、良い物だけが次世代に受け継がれる。このように考えればトータルカルチャーとサブカルチャーは相互補完的なものであり、どちらがより大きくてすぐれているか、という次元の問題とは言えない。また従来、メインカルチャーもしくはトータルカルチャーと見なされてきたようなものでも、一部のコミュニティのメンバーに閉じた文化はサブカルチャーと呼ばれるべきであり、これまでサブカルチャーの地位に甘んじていた文化であっても、もはや特定のコミュニティの専有物ではないものについてはメインカルチャーもしくはトータルカルチャーと呼ぶべきである。同様の理由で、これまでは宗教や文学などはメインカルチャーであり、ゲームやアニメはサブカルチャーであるというように十把一からげに分類されてきたが、いかにも乱暴で恣意的な方法論であって、ウィキペディアの編集履歴等をなんらかの方法で観測することによって、サブカルチャーかそうでないかをより細かく個別に、定量的かつ動的に分類できるようになるのではなかろうか。いずれにしてもこれまでは数理的な論拠もなく、これはメインカルチャーでこれはサブカルチャーであるというレッテル貼りが横行してきたのであって、ここにおいてその点に若干の異議を申し立てたいと思う次第である。

5.2 ユーザ情報の隠蔽について

Web サービスの顧客情報というものは通常は企業の最高機密であり、かつ独自の資産であって、当然厳重に管理されており、部外者が入手し得るものではない。たとえばアマゾン・コムは膨大な顧客の購買履歴情報を持っている。これをそのまま Web サービスに公開して利用することは、個人情報であるからできない。従って推薦システムというフィルタを通してサービス化し、そこからさらなる収益を得ている。別の言い方をすれば、非公開データの公開データ化、商品化とも言えるだろう。

ウィキペディアの編集履歴は、部外者が入手可能な顧客情報としては最大規模である。それは万人に完全に公開されている。しかしながら本稿で提案したシステムではそれを wikiscanner などのように暴露するために使うのではない。ユーザ情報は多くの web サービスのように隠蔽するやり方をとった。もしそれらを白日の下にさらしたければ筆者の手法を参考にして自分で作れば良いと思う。

実験をし、論文を書くときに蒐集する被験者のデータ

というものは、なかなか得難いものであり、得られてもせいぜい数十人である。大手新聞社やテレビ局ならば数千人規模のアンケートを採ることも可能かもしれないが、一個人の研究者であれば被験者が数名という例も少なくない。だが、ウィキペディアのデータベースには数十万人のユーザの編集履歴が残されており、被験者数としては圧倒的に多い。被験者が多いということは統計的により精度の高い、恣意性を排除した結果が得られるということである。またそこからまだいろいろな尺度でいろいろな結果を導き出すことができるだろうと思う。今後も、もう少しそのような可能性を模索してみたい。

5.3 ログ落ち項目の対応について

すでに何度か述べたウィキペディア日本語版特有の問題だが、編集回数が5000回を越える前に記事が一旦ログ落ちして履歴が分離される。MediaWiki の仕様上、編集履歴が長すぎると、特定版の削除を行ったときにシステムが不安定になる（クエリのレスポンスが非常に遅くなる）ために、編集回数が5000回をめどに制限をかけているという。しかし英語版などは5000回を越えていても平気で同じ記事のまま編集が続く。そのため英語版、ドイツ語版、フランス語版には George W. Bush のように何万回も編集された記事がざらにある。日本語版の現在の仕様は履歴の解析という目的には極めて迷惑な仕様だと言わねばならないが、日本語版の仕様変更を待っていてはいつまでかかるかわからないので、現行記事とログ分離記事を自動的に統合するようにしたいと考えている。

5.4 ユーザのカテゴリライズについて

トピックのカテゴリライズと双対の関係にあると考えられる、ユーザのカテゴリライズについても検討したい。推薦システムの抽出結果には、何らかの傾向やカテゴリーが反映されているように見えるが、そのカテゴリーがなんであるか、あるいは何種類のカテゴリーが関与しているかを知るには相当の推理力が必要である。しかもその推理には、カテゴリーの背後にある、どのような嗜好をもつユーザグループが、どの程度関与しているか、ということも併せて考慮する必要がある。ここで問題を整理してみると、トピックのカテゴリライズとユーザのカテゴリライズとは相互に依存していて、その双対性を利用して解析すべきだということである。また、ユーザのカテゴリそれ自体にもさまざまな応用が考えられるので、その可能性を模索しつつ開発していくつもりである。

文献

[Assouline 08] Pierre Assouline, B'eatrice Roman - Amat, Delphine

ウィキペディアンのお勧め

電車 完全一致 | 部分一致 実行

何か単語を入力します(例:「東京都」「源義経」「東京工芸大学」など)。ウィキペディア日本語版に該当する項目があれば、その編集回数と、編集を行なった登録ユーザ数などを表示します。

ここは何
ユーザ多い300
IP多い300
編集回数多い300
Wela Wiki Web(多言語版)

国鉄485系電車(485 series、일본국유철도 485계 전동차、日本國鐵485系電力動車組)⁽⁰⁾
鉄485系電車 (こくてつ485けいでんしゃ) は、日本国有鉄道 (国鉄) が設計・製造した交流流特急形車両。カテゴリ: 日本国有鉄道の高性能電車、東日本旅客鉄道の電車、西日本旅客鉄道の電車、九州旅客鉄道の電車。 (see also: [[国鉄181系電車#1000・1050・1100番台、国鉄181系電車の項]], [[国鉄183系電車#他系列からの改造車、国鉄183系電車の項]], [[国鉄183系電車#他形式からの改造車、[[国鉄113系電車#サロ110形 特急形改造車、国鉄113系電車の項]], [[国鉄24系客車]]]) テンプレート: 目撃情報、鉄道車両、vertical、Double、Vertical_images_list、main、予定、脚注ヘルプ、reflist、国鉄の交流流両用・交流電車リスト、Commonscat。

2317回編集されています。IPユーザ編集率は78.2%です。
ウィキペディアで国鉄485系電車を編集した130人の登録ユーザは以下のページも編集しています。

52人 国鉄113系電車。
50人 国鉄183系電車。
48人 国鉄103系電車。
47人 国鉄205系電車、国鉄115系電車。
44人 国鉄201系電車。
42人 ジョイフルトレイン。
41人 JR東日本209系電車。
40人 JR東日本E231系電車。
39人 国鉄165系電車。
38人 国鉄211系電車、JR西日本223系電車。
37人 国鉄キハ40系気動車(2代)、国鉄457系電車。
36人 国鉄キハ58系気動車。
35人 国鉄415系電車。
34人 鉄道事故、新幹線500系電車、新幹線0系電車、東京駅。
33人 JR西日本207系電車。
32人 国鉄EF65形電気機関車、国鉄101系電車、JR西日本221系電車、国鉄50系客車。

図3 ウィキペディアンのお勧め

- Soulas, Florence O'Kelly, Pierre Gourdain, 佐々木 勉 (翻訳) 「ウィキペディア革命～そこで何がおきているのか」岩波書店 (2008/07)
- [Bri06] Fatally Flawed - Refuting the recent study on encyclopedic accuracy by the journal Nature Encyclopadia Britannica, Inc. March 2006 (http://corporate.britannica.com/britannica_nature_response.pdf)
- [chap] Wikimedia chapters (http://meta.wikimedia.org/wiki/Wikimedia_chapters)
- [Cit] Citizendium (http://en.citizendium.org/)
- [Ehrman 06] Bart D. Ehrman, 松田 和也 (翻訳) 「捏造された聖書」柏書房 (2006/05)
- [Goo] Goo Wikipedia (http://wpedia.goo.ne.jp/)
- [hanja] Korean Hanja Wiki (http://kore.wikia.com/)
- [Knol] Knol (http://knol.google.com/)
- [klngr] The Klingon Encyclopedia (http://klingon.wikia.com/)
- [Medw] Mediawiki (http://www.mediawiki.org/)
- [Meda] 東京工芸大学芸術学部メディアアート表現学科 Wiki (http://wiki.media.t-kougei.ac.jp/)
- [Mack 05] Burton L. Mack, 秦 剛平 (翻訳) 「失われた福音書—Q 資料と新しいイエス像」青土社 (2005/03)
- [Nat 05] Special Report Internet encyclopaedias go head to head; Nature Online 14 Dec 2005 (http://www.nature.com/nature/journal/v438/n7070/full/438900a.html)
- [Ortega 09] Felipe Ortega: "Wikipedia: A Quantitative Analysis", Ph.D Thesis at Universidad Rey Juan Carlos, on April 1st, 2009 (http://libresoft.es/Members/fjlope/phd-thesis).

- [Peel 09] Michael Peel: “Wikipedia UK ‘not dying’ despite editors exodus”, Belfast Telegraph 2009/11/27
(<http://www.belfasttelegraph.co.uk/lifestyle/technology-gadgets/wikipedia-uk-not-dying-despite-editors-exodus-14578176.html>)
- [Puk] PukiWiki (<http://pukiwiki.sourceforge.jp/>)
- [reqhanja] Requests for new languages/Wikipedia Hanja
(http://meta.wikimedia.org/wiki/Requests_for_new_languages/Wikipedia_Hanja)
- [Sim] Simple API <http://www.simpleapi.net/>
- [Unc] アンサイクロペディア (<http://ja.uncyclopedia.info/>)
- [Wikt] Wiktionary (<http://ja.wiktionary.org/>)
- [Wikn] WikiNews (<http://ja.wikinews.org/>)
- [Wikia] Wikia (<http://ja.wikia.com/>)
- [Wpw] <http://www.media.t-kougei.ac.jp/wpw/>
- [Wikr] Wikirage <http://www.wikirage.com/>
- [Wikc] Wikichecker (<http://ja.wikichecker.com/>)
- [WSJ 09] “Volunteers Log Off as Wikipedia Ages”, The Wall Street Journal Online, 2009/11/27
(http://online.wsj.com/article/SB125893981183759969.html?mod=WSJ_hpp_LEFTTopStories&mg=com-wsj).
- [You] Yourpedia (<http://ja.yourpedia.org/>)
- [Zeit] ウィキペディア 多言語統計 日本語版
(<http://stats.wikimedia.org/JA/TablesWikipediaJA.htm#zeitgeist>)
- [Dump] (http://meta.wikimedia.org/wiki/Data_dumps)
- [時実07] 時実象一「ウィキペディア安易な引用はやめよう」2007-7-24朝日新聞
- [木村08] 日経サイエンス2008年12月号 p. 100掲載の記事（茂木健一郎と愉しむ科学のクオリア31「日本のネット文化を変えるには」ゲスト木村忠正（東京大学））
- [山形09] 山形浩生「ニコニコ動画のお粗末」月刊 Voice 2009年10月号 pp. 142-143
- [大塚04] 大塚英志「サブカルチャー文学論」朝日新聞社（2004/02）。
- [山本・古田08] 山本 まさき、古田 雄介「ウィキペディアで何が起きているのか—変わり始めるソーシャルメディア信仰」オーム社（2008/09）
- [池田07] 池田信夫 blog 「2ちゃんねる化するウィキペディア」（2007-12-27）、「ウィキペディアの脱構築」（2007-03-21）、「ウィキペディアのガバナンス」（2007-03-20）、「ウィキペディアとの闘い」（2007-03-15）、「ウィキペディアの嘘」（2007-03-01）（<http://blog.goo.ne.jp/ikedanobuo/>）
- [加藤99] 加藤 隆「新約聖書はなぜギリシア語で書かれたか」大修館書店（1999/04）
- [特定版削除] Wikipedia:削除依頼
(<http://ja.wikipedia.org/wiki/Wikipedia:削除依頼>)
- [著作権侵害] Wikipedia:著作権侵害への対処
(<http://ja.wikipedia.org/wiki/Wikipedia:著作権侵害への対処>)