

ウィキペディア記事閲覧回数の特徴分析

Analysis of Page Views of Wikipedia articles

曾根 広哲¹ 山名 早人^{2,3}

Hiroaki Sone¹, and Hayato Yamana^{2,3}

¹早稲田大学大学院基幹理工学研究科

¹Graduate School of Fundamental Science and Engineering, Waseda University

²早稲田大学理工学術院

² Faculty of Science and Engineering, Waseda University

³国立情報学研究所

³National Institute of Informatics

Abstract: Most of conventional Wikipedia related researches focused on various analysis of Wikipedia by using Wikipedia's database dump and its edit history. However, currently we are able to use Wikipedia's page view data in addition to use its database dump and edit history. In this paper, we focus on Wikipedia's page view data to analyze Wikipedia users' interest, i.e. page view activities, that we cannot extract only from database dump. We show the comparison of page view data both with the number of edits and the number of search results from a search engine.

1. はじめに

昨今、ウィキペディアは誰もが自由に記事を編集できる Web ベースの百科事典として一般に広く用いられている。こうした普及を背景にウィキペディアを題材にした研究も数多く行われており、その分野は多岐に渡る。ウィキペディアに関する研究が盛んに行われている理由のひとつとして、ウィキペディアの運営団体であるウィキメディア財団が編集履歴など、ウィキペディア内部で用いているデータベースのダンプデータを公開している¹ため、研究者が容易に膨大な編集履歴のデータを利用できることが挙げられる。実際に編集履歴を用いた研究は数多く存在する。

一方、ダンプデータの中には含まれていないが、ウィキペディアの各記事に対する閲覧回数も取得可能となっている²。これはウィキメディア財団の理事のひとりである Domas Mituzas 氏がウィキペディアの Squid クラスタへのアクセス統計データを公開しているものである（以下、閲覧回数データと呼ぶ）。現在のところ、この閲覧回数データを用いた研究は [1] など限定的である。

ウィキペディア研究の主なトピックのひとつとして、ウィキペディアの記事の信頼性に関する議論があり、編集履歴から信頼性を解析しようという試みが多くなされている。一方で、閲覧回数も信頼性に大きく関係すると考えられる。なぜなら、同程度に信頼性の低い記事であっても、より多く見られている記事の方が人々に誤った情報を与える可能性が高いと考えられるからである。あるいはよく見られていて、かつ編集されていない記事は多くのユーザが問題ないと判断したと考えれば、信頼性の高い記事であるとも考えることもできる。閲覧回数データは全世界のウィキペディアユーザの興味や閲覧行動を知る上で有用なデータであると考えられる。

このように閲覧回数データはウィキペディア研究において、様々な活用が考えられる。そこで、本稿では閲覧回数データの活用方法の可能性を模索することを目的として、各種の特徴分析を試みた。

2. 関連研究

2.1. 閲覧回数可視化ツール

閲覧回数データの活用の実例として、可視化ツールである Trending Topics³, Wikipedia article traffic

¹ <http://download.wikimedia.org/>

² <http://dammit.lt/wikistats/>

³ <http://www.trendingtopics.org/>

statistics⁴や Wikirank⁵などが挙げられる。これらの Web サイトでは各言語 (Trending Topics, Wikirank は英語のみ対応)でのウィキペディアの記事の人気度が一目で分かるようになっている。

2.2. 記事の評価

[1]では、閲覧回数に基づき編集者ごとの記事への貢献度、荒らしの影響などを算出する手法を提案している。Reid Priedhorsky らはウィキメディア財団から提供されたサーバへのリクエストログなどのデータから、ウィキペディア英語版の記事単位での閲覧回数を計算し、人々が見ている記事はそのほとんどがごく少数のアクティブな編集者によって書かれたものであるということを示した。また、同様に、荒らしである版が受けた閲覧回数を用いることで、ウィキペディアを閲覧したときに荒らしに遭遇する確率は低い、上昇しつつあることを明らかにした。

従来の記事の評価に関する研究では、[2]や[3]など、編集履歴に基づいて記事の信頼性を導出するものが多いが、[1]では編集履歴に加えて閲覧回数をデータとして用いている点が他の研究と異なっており、それによって新たな知見を得ている例であるといえる。

3. データセット

公開されている閲覧回数データには pagecounts ファイルと projectcounts ファイルがあり、それぞれ pagecounts-20091016-020000.gz や projectcounts-20091016-020000 のように 1 時間ごとのファイルにまとめられている。また、2009 年 10 月 29 日現在、2009 年 6 月以降のデータが取得可能となっており、最新のデータはほぼリアルタイムに更新されている。なお、ウィキペディアのリダイレクト機能については、リダイレクト元とリダイレクト先の記事双方の閲覧回数に 1 回ずつカウントされる。

3.1. pagecounts ファイル

pagecounts ファイルには 1 時間単位での記事ごとの閲覧回数が格納されており、ファイルの各行は「言語コード 記事名 閲覧回数 バイト数」となっている。例として pagecounts-20090207-140000.gz の一部を以下に挙げる。

```
en Microsoft 235 24199774
en Microsoft%20Authenticcode 1 418
en Microsoft%20Excel 1 366
en Microsoft%20Exchange 1 414
en Microsoft%20Office 1 412
```

3.2. projectcounts ファイル

projectcounts ファイルには言語ごとの 1 時間単位での閲覧回数が格納されており、ファイルの各行は「言語コード - 閲覧回数 バイト数」となっている。例として projectcounts-20091016-020000 の一部を以下に挙げる。

```
ja - 1387047 36498321434
ja.b - 2491 30321900
ja.d - 4734 50522428
ja.n - 1399 11347343
ja.q - 503 3801880
ja.s - 1038 12233207
ja.v - 257 2038036
```

なお、pagecounts ファイル、projectcounts ファイルともに言語コードの後ろに b, d, n, q などのアルファベットが付与されている。これらはウィキブックス、ウィクショナリーなどのウィキメディア財団によるウィキペディア以外のプロジェクトを指しているものと推測される。このため、pagecounts ファイルの中で「ja タイトル名～」となっているデータのみを使用し、対象を本稿で対象とするウィキペディアの日本語版に絞った。以下では特に明記のない限り、ウィキペディアとは日本語版を指すものとする。

4. 実験手法

4.1. 編集回数との比較

[1]によれば、閲覧頻度と編集頻度の間に相関はないとしている。これを確認するために、記事の編集回数や編集ユニークユーザ数と閲覧回数の比較を試みた。

⁴ <http://stats.grok.se/>

⁵ <http://wikirank.com/>

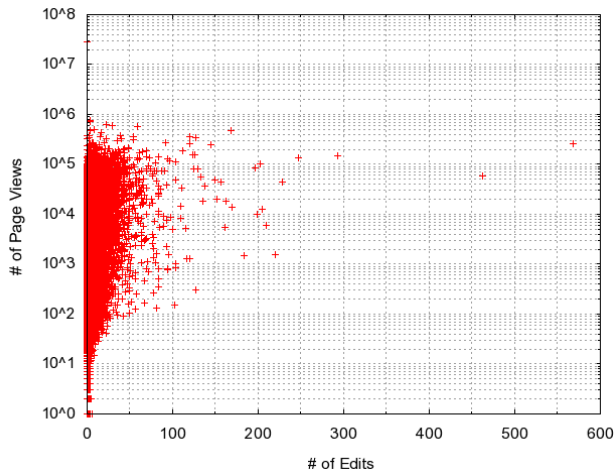


図 1: 編集回数 (2009 年 8 月) と閲覧回数 (2009 年 9 月)

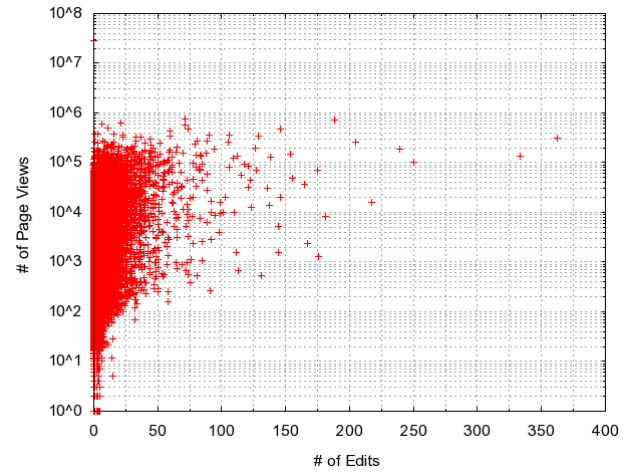


図 2: 編集回数 (2009 年 9 月) と閲覧回数 (2009 年 9 月)

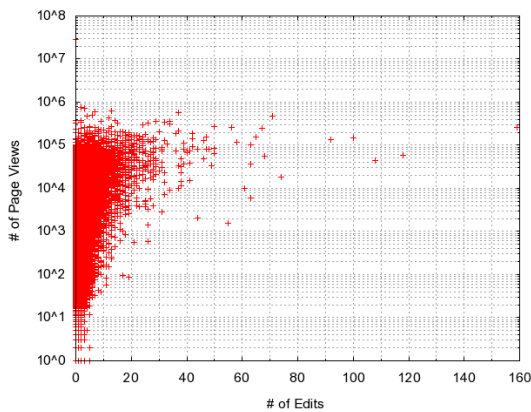


図 3: 編集ユニークユーザ数 (2009 年 8 月) と閲覧回数 (2009 年 9 月)

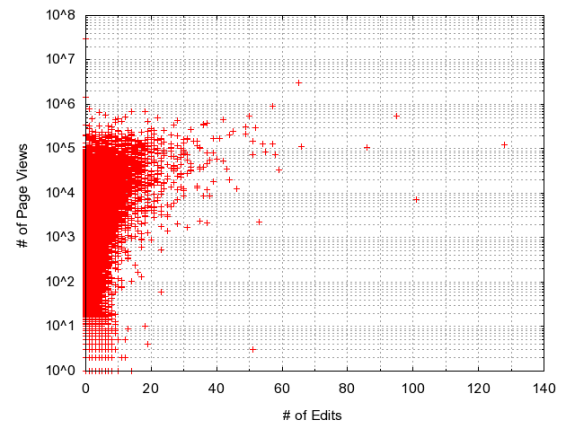


図 4: 編集ユニークユーザ数 (2009 年 9 月) と閲覧回数 (2009 年 9 月)

本実験では編集回数や編集ユニークユーザ数は 2009 年 8 月, 9 月それぞれの 1 ヶ月間のデータを用い, 2009 年 9 月の閲覧回数と比較した.

編集回数や編集ユニークユーザ数はダンプデータ `jawiki-20091019-pages-meta-history.xml.bz2` をダウンロードして用い, そこから抽出した. また, 閲覧回数の取得には 2009 年 9 月 1 日 0 時から 2009 年 9 月 30 日 23 時 0 分までの `pagecounts` ファイルを用い, Java のプログラムによって 2009 年 9 月の各記事の閲覧回数の総和を求めた.

4.2. 検索エンジンのヒット数との比較

ウィキペディアの閲覧頻度は世間の関心や語の有名度を示していると考えられる. 一方, 論文などでは語の有名度を測るための指標として検索エンジンのヒット数が用いられている. そこで, 検索エンジンのヒット数と閲覧回数の比較を試みた.

検索エンジンには日本で最大のシェアを持つ Yahoo! JAPAN を利用し, 検索 API⁶を通じて検索結

果の取得を行った. 取得期間は 2009 年 4 月 15 日~5 月 3 日で, クエリとしてウィキペディアの全記事名 (2009 年 1 月時点 全 562,579 本) をひとつずつ用い, ヒット数を取得した. その際のオプション設定の値を表 1 に示す.

表 1: 検索 API のオプション設定

言語設定	日本語
似たページ	許可
アダルトフィルタ	オフ
フレーズ検索	使用せず

閲覧回数の取得方法は 4.1 節と同様であるが, 2009 年 4 月 1 日 0 時から 2009 年 4 月 30 日 23 時 0 分までの `pagecounts` ファイルを用い, 2009 年 4 月の閲覧回数を求めた.

5. 実験結果

5.1. 編集回数との比較

実験結果を図 1, 図 2 に示す. 図 1 は横軸が 2009 年 8 月の編集回数, 縦軸が 2009 年 9 月の閲覧回数と

⁶ <http://developer.yahoo.co.jp/webapi/search/>

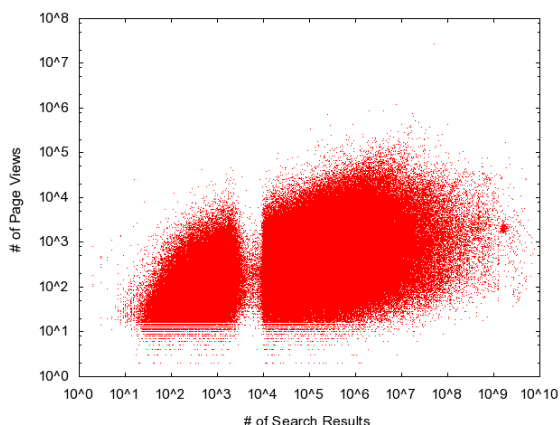


図5: 検索エンジンのヒット数と閲覧回数

なっており、図2は横軸が2009年9月の編集回数、縦軸が同じく2009年9月の閲覧回数である。

また、編集ユニークユーザ数と閲覧回数との関係を図3、図4に示す。図3は2009年8月の編集ユニークユーザ数と2009年9月の閲覧回数であり、図4は2009年9月の編集ユニークユーザ数と2009年9月の閲覧回数との比較である。

これらの結果より、編集回数や編集ユニークユーザ数と閲覧回数との間に明らかに相関は見られず、人々によく見られる記事とよく編集される記事は異なるということが確認された。従って、研究対象のデータとして閲覧回数を用いることで、編集履歴のみからでは知り得ない一般の閲覧ユーザの興味を捉えることができるようになると思われる。

5.2. 検索エンジンのヒット数との比較

実験結果を図5に示す。また、Spearmanの順位相

関係数を求めたところ、0.53となった。つまり、ヒット数が高いものは閲覧回数も大きくなる傾向がある。

また、ヒット数10,000件の付近は他と比べて明らかに挙動が異なっている。そこで、より詳しく調べるためにヒット数についてヒストグラムにしたものが図6である。図を見ると、やはり 10^3 件から 10^4 件の間のみ挙動が異なっている。これらのことから、Yahoo! JAPANでのヒット数算出のアルゴリズムが 10^3 件から 10^4 件の間でのみ異なっているのではないかと推測される。

このようなヒット数の挙動からも分かるように、ヒット数は検索エンジンの内部でどのような処理によって算出されているかは不明である[4]。しかし、ウィキペディアの閲覧回数はサーバのアクセスログという数値の出所が明らかにされており、かつヒット数との相関を持つので、より良い語の有名度についての指標としての活用が考えられる。

6. おわりに

本稿では、ウィキペディアの閲覧回数の特徴分析を目的として、実験を行い、その活用方法を模索した。その結果、閲覧回数は編集回数や編集ユニークユーザ数との相関は見られず、よく編集される記事とよく見られる記事は異なることや検索エンジンのヒット数との相関は見られることが確認できた。閲覧回数を用いることで編集回数のみからでは得られないユーザの興味を捉えることができるようになるのではないかと考えられる。また、語の有名度を測る指標のひとつとしてなど、様々な活用方法があるのではないかと考えている。

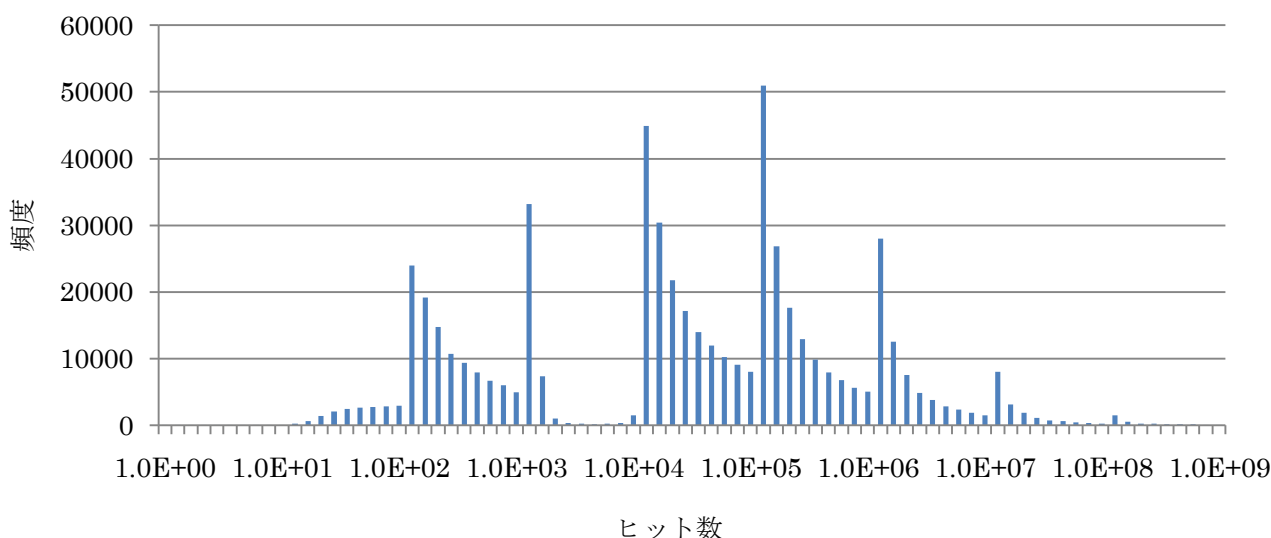


図6: ヒット数のヒストグラム

参考文献

- [1] Reid Priedhorsky, Jilin Chen, Shyong (Tony) K.Lam, Katherine Panciera, Loren Terveen, and John Riedl. Creating, destroying, and restoring value in wikipedia. Association for Computing Machinery GROUP '07 conference proceedings, 2007.
- [2] B. Thomas Adler, Luca de Alfaro and Ian Pye. Measuring Author Contributions to the Wikipedia. WikiSym 2008, 2008.
- [3] T. Cross. Puppy smoothies: improving the reliability of open, collaborative wikis, 2006.
- [4] 舟橋卓也, 上田高德, 平手勇宇, 山名早人. 商用検索エンジンのヒット数に対する信頼性の検証. 日本データベース学会論文誌, Vol.7, No.3, pp31-36, 2008.