査読付論文

Wikipediaにおける知的生産活動の構造とプロセスに関する研究

大野　晋・稲葉　光行

要旨

本研究の目的は、不特定多数のアカウントによるウェブ上での協調的な知的生産活動の構造とプロセスを理解することである。本論文では、その成功例の一つとされるWikipediaを取り上げ、その中でも特に、模範的な記事とされる「秀逸な記事」、「良質な記事」、および「おすすめ記事」と呼ばれる記事群を対象とし、それらにおける登録アカウントと匿名アカウントの編集行為の相違についての分析を行った。また我々は、記事編集に関するメタ的な議論の場である「ノートページ」の分析を行った。その結果、匿名アカウントによる記事編集回数の総計が、登録アカウントの編集回数の総計を上回っていること、また、直接的な編集ではなく、議論を介して間接的に記事の拡充に貢献するアカウントが多数存在することが示された。

Ⅰ．研究の目的

本研究の目的は、ウェブを通して行われる知的生産活動に注目し、その活動の構造とプロセスを理解することである。本研究では、このような知的生産活動に対し、科学計量学の視点から分析を行う。科学計量学とは、知的生産活動に関する数量的なデータを収集し、分析、考察を行う研究領域である。本論文の研究のポイントは以下の3つである。

1）ウェブ上での知的生産活動の成功例の一つとして注目されるWikipediaにおける編集行為に関する分析・考察を行なう。
2）Wikipediaの記事の中でも、特に模範的とされる記事群に焦点をあて、それらの記事編集に誰がどのように関わっているのかを分析する。具体的には、記事本体とノートページに対する編集行為を、登録アカウントおよび匿名アカウントそれぞれの集計データを元に考察する。
3）一般的な知的生産活動に関する古典的研究として「Lotkaの法則」（Lotka, 1926）を取り上げ、その応用研究として、Wikipediaのようなウェブ上での集合知構築に、Lotkaが見出した「不均衡性」が成立しているかどうかを分析する。

Ⅱ．関連研究

本章ではまず、本研究における分析のベースであるLotkaの法則の概要について述べる。次に、Lotkaの法則に関わる研究事例を紹介する。

Ⅱ.1　Lotkaの法則

Lotkaの法則（Lotka, 1926）は、知的生産活動を計る際に用いられる主要な指標の1つである。Lotkaは、科学者の論文生産を知的生産活動と見立て、科学者とその論文の量の割合を調査した。その結果、生産者と生産物の量に関する不均衡が存在することを指摘した。Lotkaの法則は以下の式で表される。

$$X^nY = C$$

もしくは

$$Y = C/X^n$$

Xは、科学者が生産する論文の数、Yは論文数Xを生産する科学者の人数の度数（相対的割合）であり、Cは任意の定数である。対象となる知的生産活動を数量化した結果、nの値が2に近似する場合は、その活動がLotkaの法則に適合するとされる。

Lotkaは、1907年から1916年において、論文誌の調査を行った。Lotkaによる調査では、6,891人の科学者のうち上位84人が、論文総数22,839本のうち4,200本を生み出しているという結果が得られた。そして、科学者と論文生産数の間に、逆二乗の法則が成立しているこ

とを見出した。つまり彼は、論文生産という活動において、高い生産性を示す科学者の割合は小さく、低い生産性を示す科学者の割合は大きくなるという法則が成立する可能性を指摘した。

Lotka の分析は、化学と物理における科学者とその論文の生産数に関するものであったが、その後、他の領域の論文生産活動においても、Lotka の法則が成立する可能性が示された。例えば Price（1963）の調査では、科学者の論文の生産性において上位 6% の科学者が 50% の論文を生産しているという結果が得られている。

さらに、Lotka の法則は、論文生産以外の知的生産活動を分析する際の指針として用いられている。例えば Newby（2003）は、オープンソースソフトウェアにおける開発者とソフトウェアの本数の関係を、Lotka の法則の視点から調査した。具体的には、Linux で使用できるソフトウェアのデータベースサイトである LSM（Linux Software Map）と、オープンソースソフトウェア開発のためのリポジトリサイトである SourceForge を対象とし、登録された開発者数とそのソフトウェア数との関係を、Lotka の法則における科学者と論文数の関係に対応させた分析を行った。Lotka による調査と異なるのは、Lotka が論文の筆頭著者のみを扱っているのに対し、Newby らは、ソフトウェアの制作に関わったすべての開発者を対象としている点である。その結果、ソフトウェアの開発者と本数との関係が、Lotka の法則で示された逆二乗には至らなかったものの、少数の開発者が多数のソフトウェアを登録しているという点においては Lotka の法則を支持するデータが得られた。

本研究と同様に、Wikipedia における知的生産活動を Lotka の法則の視点から考察したものとしては、Voss（2005）による Wikipedia の調査が挙げられる。Voss は、2004 年のドイツ語版 Wikipedia のデータを対象とし、記事の編集回数と編集者数に関して、Lotka の法則への適合性を調査した。結果として、n の値が 1.5 となり、逆二乗になることはなかったが、多くの記事を編集している著者がより多くの記事を編集するという傾向がある可能性が示唆された。Voss が対象としたデータは 2004 年のものであり、当時の Wikipedia 日本語版の記事数は 10 万件に満たなかったが、2009 年時点での日本語版の総記事数は約 62 万件であり、6 倍以上に増加している。従って本研究では、最新のデータセットを用いて、Lotka の法則への適合性を改めて調査する。

Wikipedia を知的生産活動の視点から調査したその他の研究としては、Almedia（2007）による調査が挙げられる。Almedia は、2006 年の Wikipedia 英語版のデータを用いた調査を行った。ここでは、少数の編集者が多数の記事を執筆しており（n=1.63）、大多数の編集者はあまり記事を編集することがない（n=0.65）という結果が得られている。Almedia の研究では、「秀逸な記事」や「良質な記事」といった、記事の種類別の分析が行われていない。従って、本研究では、記事の種類にも着目し、より詳細な視点から Wikipedia における知的生産活動を分析する。

Ⅲ．データと研究方法

以下では、本研究で用いたデータと分析手法について述べる。

Ⅲ.1　元データ

Wikipedia のサイトは、MediaWiki と呼ばれるオープンソースの CMS（Contents Management System）によって構築されており、データベース管理システムとしては MySQL が採用されている。また、多少のタイムラグはあるものの、MySQL のダンプデータをインターネット経由でダウンロードし、Wikipedia をローカル環境にセットアップすることもできる。本研究では、2009 年 7 月 13 日時点における日本語版 Wikipedia の本文以外の履歴データをダンプしたファイル（jawiki-20090713-stub-meta-history.xml.gz）を取得し、ローカル環境にセットアップした後、データの集計と分析を行なった。

Ⅲ.2　記事のデータ

本研究では、特に Wikipedia における次の三種類の記事を対象とした集計と分析を行なった。

1）「秀逸な記事」：Wikipedia の百科事典としての価値を高めることを目的とし、Wikipedia のユーザ投票によって選ばれる記事（85 件）
2）「良質な記事」：高い質を保ち、「秀逸な記事」に近い記事。Wikipedia のユーザによって査読が行われ選ばれる記事（370 件）
3）「おすすめ記事」：自薦他薦によって、「秀逸な記事」やその選考に勧めたい記事（279 件）

本論では、上記の三種類の記事群の総称を、「模範的な記事」と呼ぶこととする。

例えば、「模範的な記事」の1つである「秀逸な記事」は、Wikipediaコミュニティによって、以下の基準を満たすものと定義される（Wikipedia, 2009a）。

1）その主題を扱う専門家（研究者、実務家、その他）から見て、百科事典において必ず説明されるべきことが全て説明されている。ただし、何が必須かは部分的には関連記事との連携・分担関係にもよる。
2）詳しくない読者にもその主題について理解できるように、わかりやすく書かれている。ただし、高度に専門的な主題を扱ったものであれば、関連記事を読んで理解していることを前提にするのは問題ない。
3）内容が充実している。必須の点だけをわかりやすくカバーしただけでは不十分。
4）完成度が高い。文章が読みやすい、構成がしっかりしている、明らかに未完成な部分がない、（可能なら）図や画像やなどがついている、など。
5）観点の中立性が保たれている。
6）「出典」または「参考文献」が挙げられている。
7）以上の点が全て満たされている。

Ⅲ.3　ノートページのデータ

Wikipediaの各記事には、ノートページと呼ばれる、記事の質を高めるためのメタ的な議論のためのページがある。ノートページの目的は以下の通りである（Wikipedia, 2009b）。

1）いろいろな立場の人の見方を取り交し、それをすり合わせて中立的な記述にする
　その記事でフォローして欲しいことについて要望する。
2）記事の内容に疑問を感じるが、書き直せるほどの知識はないので、疑問だけ提示しておく
3）内容が重複している記事を指摘し、統合を提案する

本研究では、「模範的な記事」の編集プロセスを総合的に理解するため、記事に加えて、ノートページに対する編集行為についても集計・分析を行なうこととした。

Ⅲ.4　アカウントの種別

Wikipedia上で編集を行う方法には、次の2つがある。1つは、アカウントを登録し、そのアカウントでログインし、編集を行う方法である。もう1つは、アカウント登録をせずに、匿名のままで編集を行う方法である。この場合、アカウント名に代わる情報としてIPアドレスが記録される。本研究では、前者の方法で用いられるアカウントを「登録アカウント」と呼ぶ。後者の方法で記録されるIPアドレスを、「匿名アカウント」と呼ぶ。

本研究では、「模範的な記事」の編集に関して、これらの2種類のアカウントがどのように関与しているかを集計・分析する。

「登録アカウント」の編集は、実アカウント毎の編集行為として記録される。そして「匿名アカウント」の場合は、書き込み時に用いた機器のIPアドレスがアカウント名に代わるものとして記録される。そのため、IPアドレスが動的に変化する機器からの書き込みは、同一機器でも異なるものとして記録される。さらに、プロキシサーバを導入している大学や企業内部の機器からの書き込みは、すべて同一IPアドレスから発信されたものと見なされる。

Ⅳ．分析と考察

以下では、Wikipedia上での「模範的な記事」、それぞれのノートページ、およびⅢ.4で述べたアカウント種別毎の書き込みに関する集計と、それらに対する分析・考察について述べる。

Ⅳ.1　記事とノートページの編集者数の集計

以下では、Wikipedia上で、模範的とされる記事に対する編集を行ったアカウント数の割合に関する集計結果を示す。図1は「秀逸な記事」、図2は「良質な記事」、図3は「おすすめ記事」に関する記事とノートページの編集者数の集計結果である。

図中の左側の円グラフは、記事の編集者数と、ノートページに対する編集者数の割合を示す。図中の右側の縦棒グラフは、記事およびノートページを編集したアカウント数と、ノートページのみを編集したアカウント数の割合を示している。

3種類の記事のすべてにおいて、記事のみを編集しているアカウントの割合が圧倒的に大きいが、記事とノートページに書き込みをしているアカウントが5～6%、ノートページのみに書き込むアカウントも約2%存在し

ている。Wikipedia に関しては、ネット上でのボランタリーな記事書き込みによる集合知構築が注目されることが多いが、「模範的な記事」においてさえ、記事は一切書かず、ノートページのみへの書き込みによる間接的な参加を行うアカウントが存在するという点は興味深い。

Ⅳ.2　登録アカウント数と匿名アカウント数の集計

図 4 ～ 6 は、Wikipedia 内に自らのアカウントを登録し編集を行っている「登録アカウント」と、アカウント登録をしないまま利用する「匿名アカウント」による記事の編集者数に関する集計結果を示している。

コミュニティ内での投票に基づいて選び出される「秀逸な記事」および「良質な記事」においては、編集者数における登録アカウントの割合が 30% 台、匿名アカウントの割合が 60% 台と、類似した結果となっている。

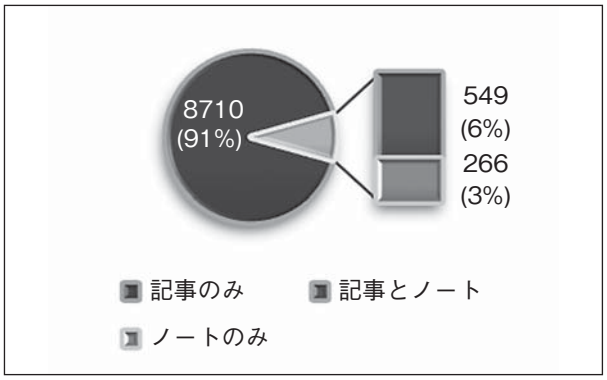

図 1：「秀逸な記事」の総編集者数における記事への編集者数とノートページへの編集者数の割合

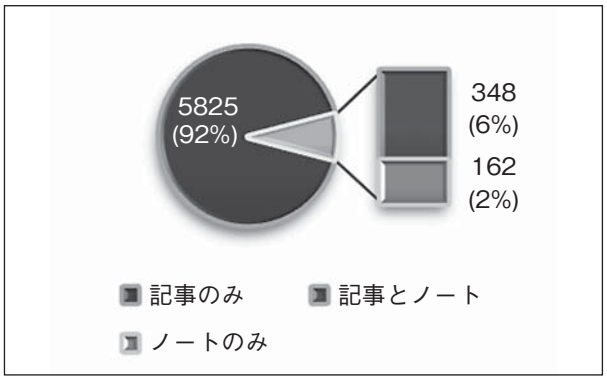

図 2：「良質な記事」の総編集者数における記事への編集者数とノートページへの編集者数の割合

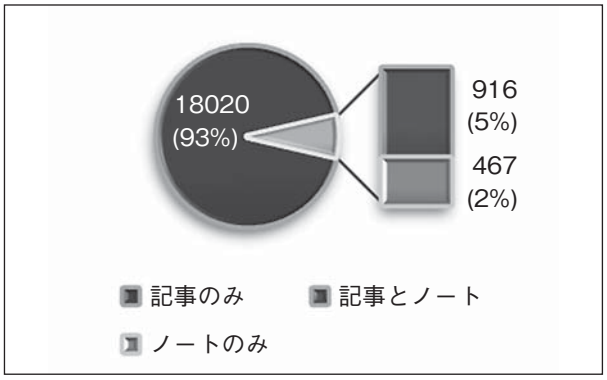

図 3：「おすすめ記事」の総編集者数における記事への編集者数とノートページへの編集者数の割合

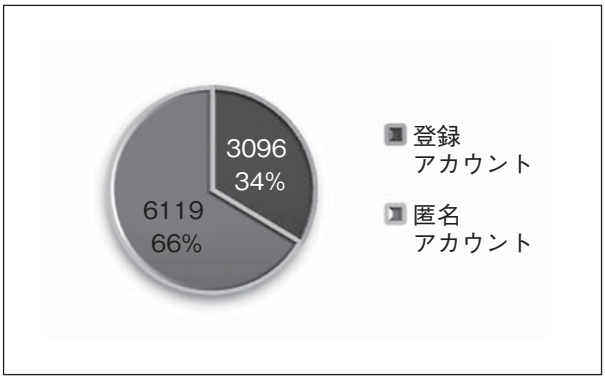

図 4：「秀逸な記事」における登録、匿名アカウントの数の割合

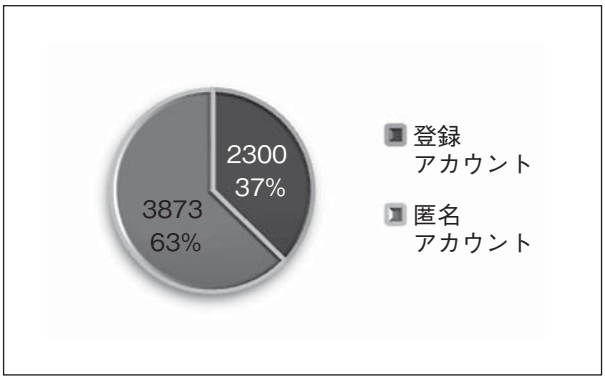

図 5：「良質な記事」における登録、匿名アカウントの数の割合

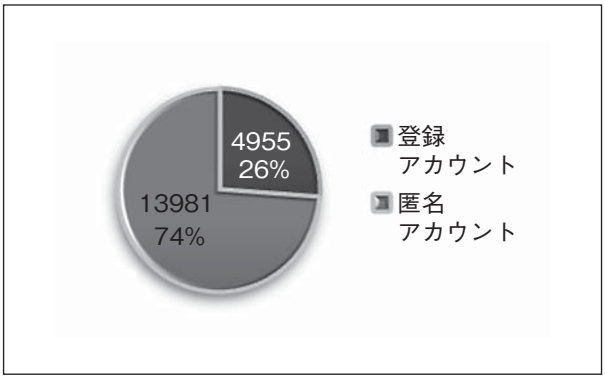

図 6：「おすすめ記事」における登録、匿名アカウントの数の割合

しかし､個々のアカウントによる自薦または他薦という、より低い「敷居」に基づいて選び出される「おすすめ記事」においては、編集回数における登録アカウントの割合が20% 台、匿名アカウントの割合が70% 台であり、匿名アカウントが占める割合が多少高くなっている。

また3種類の記事すべての編集回数においては、アカウント登録という、より積極的な関与の姿勢を示している登録アカウントよりも、アカウントを登録していない匿名アカウントの割合がより大きいことが確認された。

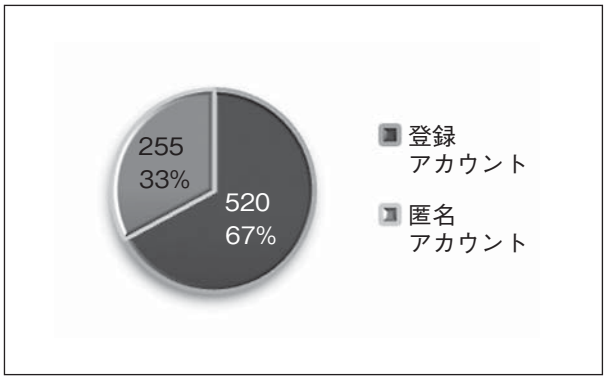

図7：「秀逸な記事」ノートページにおける登録、匿名アカウントの数の割合

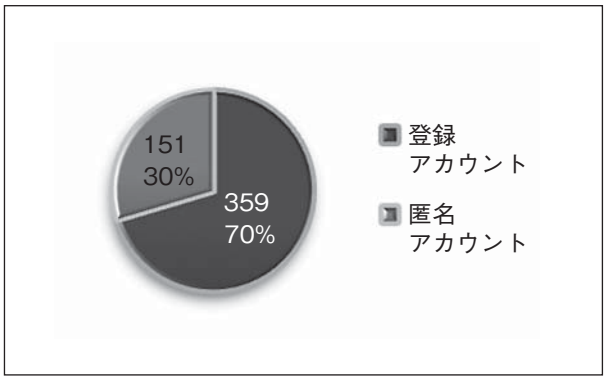

図8：「良質な記事」ノートページにおける登録、匿名アカウントの数の割合

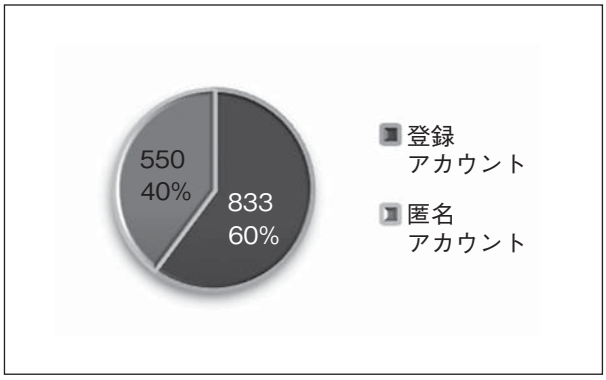

図9：「おすすめ記事」ノートページにおける登録、匿名アカウントの数の割合

図7～9は､記事に関するメタ的な議論を行うための、ノートページにおける編集回数の集計結果である。前述した3種類の記事では、匿名アカウントによる編集回数の割合が高かったが、ノートページでは、両者の編集回数の割合が逆転している。

この記事とノートページの編集回数の割合に関しては､「おすすめ記事」における編集回数の割合の解釈と同じく､「敷居」の高さが原因として考えられるが、より詳細な調査、及び考察が必要である。

Ⅳ.3　Lotkaの法則に基づく考察

以下では、編集回数の集計結果と、それらに対するLotkaの法則に基づく考察について述べる。ここでは、Wikipedia上での編集回数および編集人数を、Lotkaの式のXの値として扱い、また編集人数における編集回数の割合を、Lotkaの式におけるYの値に該当させて分析を行う。

Ⅳ.3.1　記事と登録アカウントとの関係

図10～12は､「秀逸な記事」､「良質な記事」､および「おすすめ記事」を対象とした、登録アカウントの記事の編集人数と編集回数の関係を示している。Lotkaの式に当てはめた結果、nの値は1.1から1.3になり、Lotkaの法則に適合する2の近似値にはならない。しかし、少数のアカウントが多くの記事の編集を行っているという、Lotkaの法則に近い傾向があることが観察された。

Ⅳ.3.2　記事と匿名アカウントとの関係

図13～15は､「秀逸な記事」､「良質な記事」､および「おすすめ記事」を対象とした、匿名アカウントの記事の編集人数と編集回数の関係を示している。ここでは、nの値は2の近似値となり、Lotkaの法則に合致した現象が観察された。

Ⅳ.3.3　ノートページと登録アカウントとの関係

図16～18は､「秀逸な記事」､「良質な記事」､および「おすすめ記事」のノートページを対象とした、登録アカウントの記事の編集人数と編集回数の関係を示している。ここでは、nの値は1.2から1.5であり、Lotkaの法則に適合する2の近似値が得られたとは言いがたいが、少数のアカウントが多くの記事の編集を行っているという、Lotkaの法則に近い傾向があることが観察された。

Ⅳ.3.4　ノートページと匿名アカウントとの関係

図 19 ～ 21 は、「秀逸な記事」、「良質な記事」、および「おすすめ記事」に付随するノートページを対象とした、匿名アカウントの記事の編集人数と編集回数の関係性を示している。ここでは、n の値は 2 の近似値となり、Lotka の法則に合致した現象が観察された。

Ⅳ.3.5　全体的な考察

記事、ノートページに対し、登録アカウント及び匿名アカウントが行った編集回数の割合を、Lotka の法則に基づいて分析を行った。結果、n の値にばらつきが見られたものの、Lotka の法則に近い傾向が確認された。同時に、編集回数が 1 回であるアカウント数が多いことが n の値をゆがめていることが観察された。従って、書き

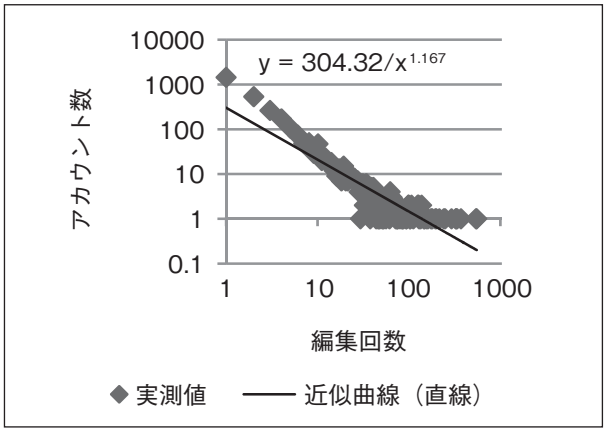

図 10：「秀逸な記事」- 登録ユーザ

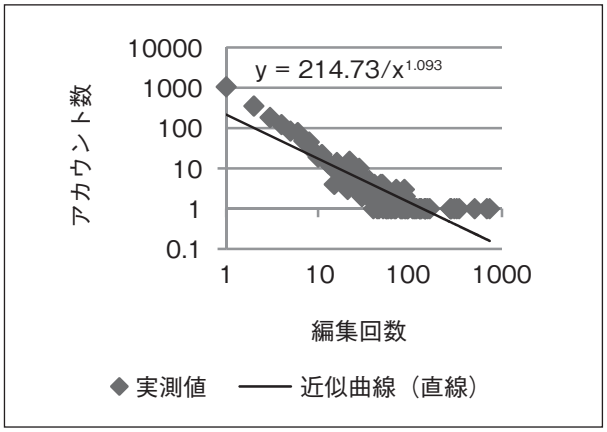

図 11：「良質な記事」- 登録ユーザ

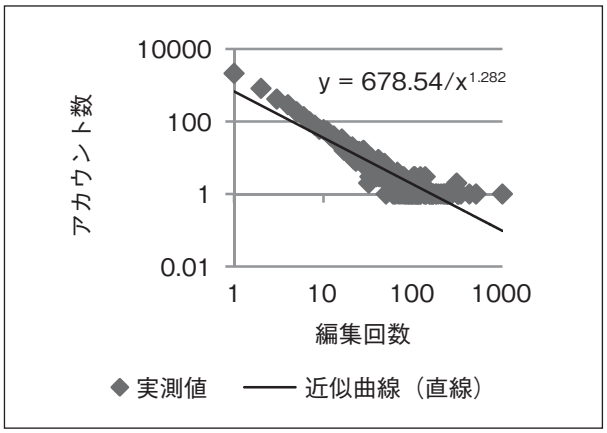

図 12：「おすすめ記事」- 登録ユーザ

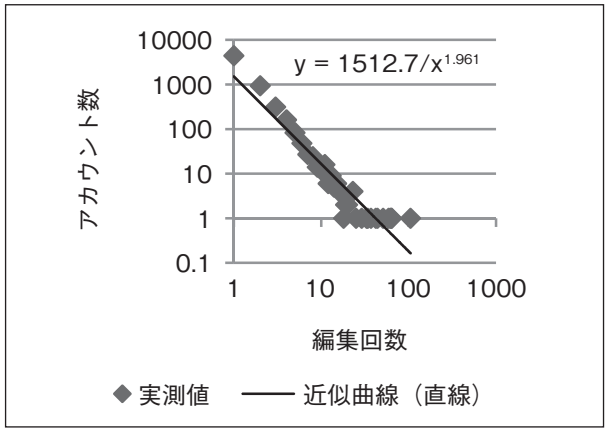

図 13：「秀逸な記事」- 匿名ユーザ

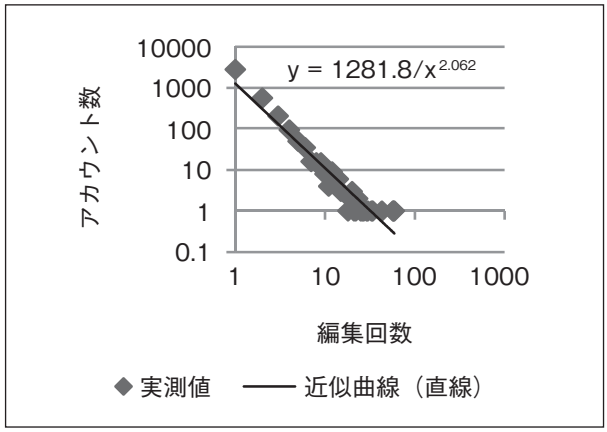

図 14：「良質な記事」- 匿名ユーザ

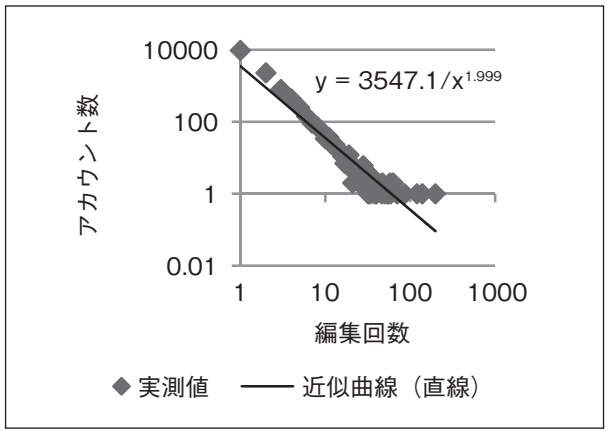

図 15：「おすすめ記事」- 匿名ユーザ

込みの多いサンプルと少ないサンプルを分割した上で分析するなどの工夫が必要であろう。

V．まとめ

本研究では、Wikipedia における記事とノートページの編集行為を対象として、ウェブ上での知的生産活動の構造とプロセスに関する分析と考察を行った。尚、本研究においては、記事に対する編集行為における構造とプロセスを一体のものとして扱った。

具体的な対象として、Wikipedia における模範的な記事である「秀逸な記事」、「良質な記事」、および「おすすめ記事」、またそれらの記事に付随するノートページに関して、編集回数の集計を行った。さらに、「登録ア

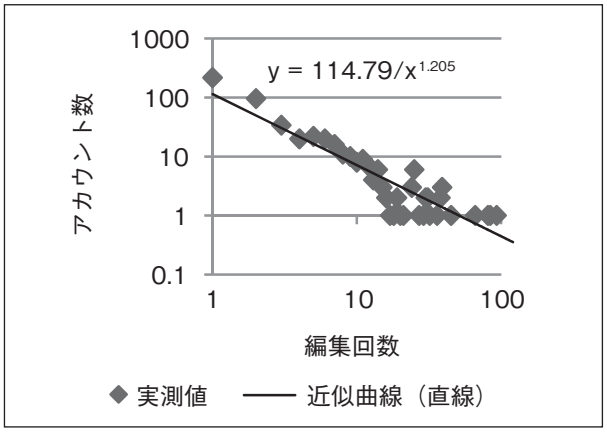

図 16：「秀逸な記事」ノートページ - 登録ユーザ

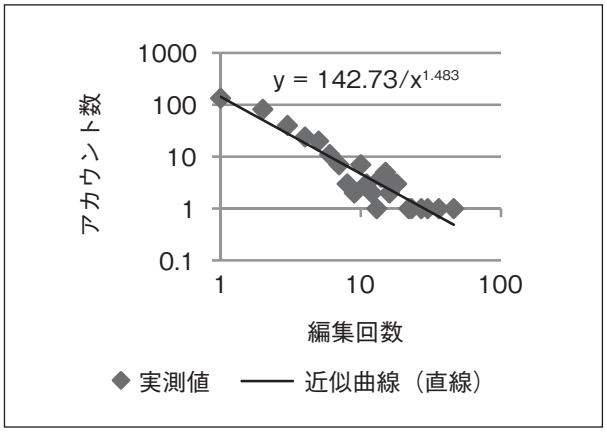

図 17：「良質な記事」ノートページ - 登録ユーザ

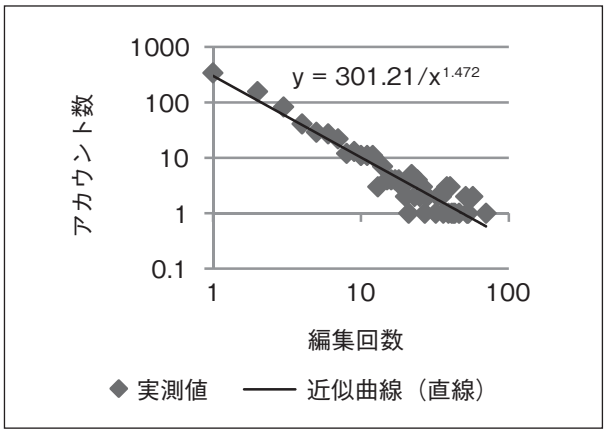

図 18：「おすすめ記事」ノートページ - 登録ユーザ

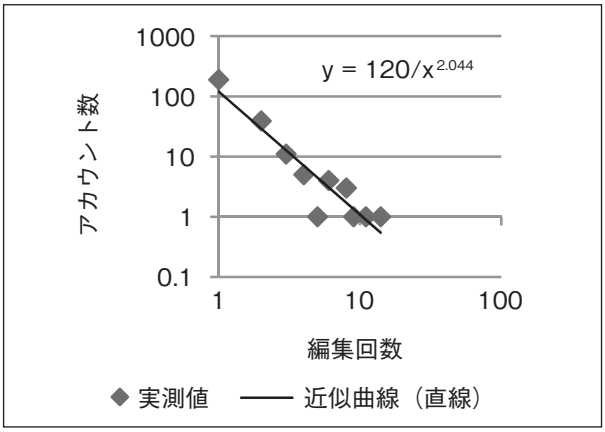

図 19：「秀逸な記事」ノートページ - 匿名ユーザ

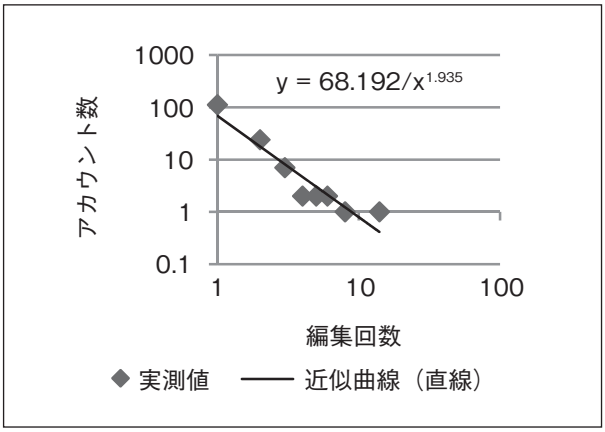

図 20：「良質な記事」ノートページ - 匿名ユーザ

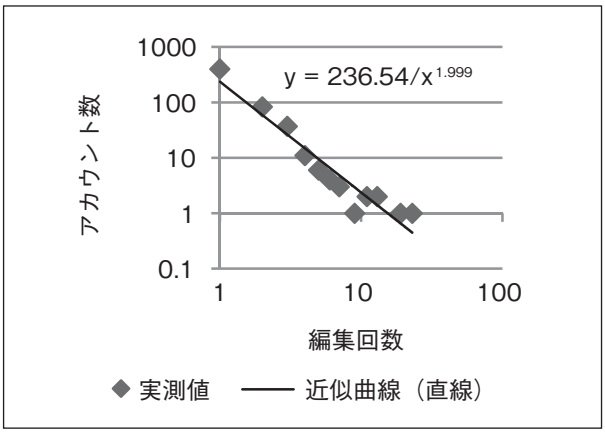

図 21：「おすすめ記事」ノートページ - 匿名ユーザ

カウント」や「匿名アカウント」のそれぞれが、3種類の記事やノートページの編集にどのように関わっているかという点を分析した。さらに、Lotkaの法則に基づき分析を行い、記事への編集行為に関する考察を行った。本研究によって得られた主な知見は以下の通りである。

1）記事とノートページの編集者数の集計によれば、記事に対する編集を行うアカウントが多いことが観察されるのは自然なことであるが、ノートページのみに編集するアカウントも一定数存在することが確認された。つまり、Wikipediaでは、議論を通して観察的に記事の拡充に貢献しようとするアカウントが多数存在することが明らかになった。
2）Wikipedia上では、3種類の記事すべての編集回数において、アカウント登録という積極的な関与の姿勢を示している登録アカウントよりも、アカウントを登録していない匿名アカウントの割合がより大きいことが確認された。
3）先行研究同様、Wikipedia上の記事編集行為のすべてが、Lotkaの法則で提案されている傾き（nの値）とは必ずしも合致していないが、全体として、少数のアカウントが多くの記事の編集を行っているという、Lotkaの法則に近い傾向があることが観察された。

本研究では、模範的とされる記事（「秀逸な記事」、「良質な記事」、および「おすすめ記事」）と、それらに関するメタ的な議論を行っているノートページに焦点を当てて分析をおこなったが、Wikipedia上での知的生産活動の全体像を把握するためには、Wikipedia上の記事一般を対象とした分析を進めていく必要がある。また本研究では、編集回数に着目し、「量的」なデータに基づく考察を行ったが、具体的にどのような編集が行われたかなどの「質的」な側面についてのデータ収集と分析は行っていない。

今後の検討課題の1つとして、量的な視点については、書き込みの多いサンプルと少ないサンプルを分割した上での分析が挙げられる。また、これらの量的な分析に加えて質的側面からのデータ分析と考察を行うことで、Wikipediaにおける知的生産活動に関する総合的な分析を進めていく必要がある。

参考文献

Almedia, R. B, Mozafari, B, and Cho, J., (2007), “On the Evolution of Wikipedia”, *ICWSM*, 2007.

Lotka, A. J. (1926), “The frequency distribution of scientific productivity”. *Journal of the Washington Academy of Sciences*, 1926.

Newby, G. B., Greenberg, J., and Jones, P. (2003), “Open source software development and Lotkas Law: Bibliometric patterns in programming”. *JASIST*, 54 (2), 2003.

Price, D. J., *Little Science, Big Science*. Columbia University Press, 1963.

Voss, J. (2005), “Measuring wikipedia”, *ISSI*, 2005.

Wikipedia: 秀逸な記事の選考 － Wikipedia. (n.d). Retrieved November 15, 2009, from http://ja.wikipedia.org/wiki/WP:FAC

Help: ノートページ － Wikipedia. (n.d). Retrieved November 15, 2009, from http://ja.wikipedia.org/wiki/H:TP