1H2-4

Wikipediaの成長パターンにおける多様性

Diversities in Wikipedia's growing patterns

山崎 由佳[*1] 伊藤 貴一[*1] 井庭 崇[*1] 熊坂 賢次[*1]
Yuka Yamazaki Takaichi Ito Takashi Iba Kenji Kumasaka

[*1] 慶應義塾大学 政策・メディア研究科
Graduate School of Media and Governance, Keio University #1

In this research, we explore rules and diversities in Wikipedia's growing patterns. Wikipedia is an encyclopedia on the Web. First, we analyzed the distributions of the frequencies of the linked words in several categories. The analysis gave a result that the distributions are raising to a high power. But, the differences between the processes of growth still remain unclear. In this research, we explore the following points; (i) common rules running through the whole categories, (ii) a huge variety of individual growing patterns. To be more precise, we analyzed the annual changes in distributions of the frequencies of the linked words in Japanese Wikipedia articles. Then, we plot the data into double logarithm graphs, and calculate coefficient of regression to formulate entire trends. Additionally, by using the regression coefficient as parameters, we make clusters on Self-Organizing Maps to lead diversities growing patterns.

1. はじめに

近年，ウェブページへのアクセス頻度や地震の規模など，さまざまな分野において，「べき乗分布」(power law distribution)を示す現象が発見され，注目を集めている． そこで，まず，WWW上で多くの人々の相互編纂によって作られるWikipedia (http://ja.wikipedia.org)においても，そのようなべき乗即が成立するかということを調べた．Wikipediaには，一般的に流通する紙ベースの辞書と同様に，記事を分類するカテゴリが設けられている．そのいくつかのカテゴリを対象とし，各カテゴリの記事内にあるハイパーリンクの被使用頻度と順位の関係を分析したのである．すると，やはりべき乗即を見て取ることができた．

しかし，たしかに，いくつかのカテゴリを対象として，ハイパーリンクの使用頻度と順位の関係におけるべき乗分布を発見したが，一方で，それらの分布は経年で同じような経緯を辿ってきたのか，という疑問が浮かび上がった．そこで，本研究では，Wikipedia の各カテゴリの記事内で使用されるハイパーリンクの使用回数と順位の分布について，2004 年から 2008 年の 5 年間のデータをもとに，その成長過程の分析を行い，そこにどのような差異があるのかということを明らかにする．

2. 分析対象

本研究では，Wikipedia の記事内で使用されるハイパーリンクをデータとして，カテゴリ別に収集した．経年のデータを使用するため，2004 年より各年の 1 月 1 日時点で表示される記事データを履歴より取得した．なお，データ収集は 2008 年 1 月に行った．

分析対象とするカテゴリは，Wikipedia で，検索の起点として定められている 9 つの「主要カテゴリ」直下にあるカテゴリとした．表 1 は，分析対象とするカテゴリ数および記事数の一覧である．分析対象は，カテゴリ数 352，記事数 19,425 となった．

連絡先：山崎由佳，慶應義塾大学 政策・メディア研究科，
藤沢市遠藤 5322 o408，yuca@sfc.keio.ac.jp

主要カテゴリ	直下カテゴリ数	記事数
科学	32	1,166
学問	76	4,951
技術	58	3,094
自然	34	2,168
社会	45	3,175
人間	17	999
総記	7	378
地理	19	835
文化	43	1,922
歴史	21	737
計	352	19,425

表 1 分析対象カテゴリ数および記事数一覧

3. 分析手法

分析手法は以下のとおりである．

3.1 べき指数の傾向の可視化

まず，ハイパーリンクの被使用頻度と順位を，両対数グラフにプロットする．次に，近似式を算出する．近似式は，以下の式で表される．

$$LogY = \beta + \alpha * LogX$$

また，算出したべき指数の評価にあたって，色づけを試みる．べき指数の値が大きいほど，色の濃さが増すように可視化する．それによって，全体の経年での傾向を可視化し，概観する．

3.2 自己組織化マップによるクラスタリング

全体の傾向を色づけによって可視化した後，個別の成長パターンを探るために，近似式の傾きを変数として，自己組織化マップを用いてクラスタリングを行う．自己組織化マップとは，競合学習を基礎とする人工ニューラルネットワークの一種であり，教師なし学習を行うものである．このアルゴリズムを用いることで，

経年で似た傾向をもつカテゴリ同士をクラスタリングし，成長パターンを描き出す．

4. 成長パターンの分析

4.1 Wikipedia のカテゴリ全体の成長の傾向

まず，Wikipediaのカテゴリ全体の傾向について分析する．分析対象となった 352 カテゴリすべてにおいて，2004 年から 2008 年の 5 年間分のハイパーリンクの被使用回数と順位の両対数グラフを作成し，近似式を算出した．そして，自己組織化マップによるクラスタリングおよび各年におけるグラフの傾きに色づけを行ったものが図 1 である．それぞれのクラスタにある 2 つの列が，左から傾きの色づけ，記事数の増加率の色づけとなっている．この 2 列は，縦に一つ一つのカテゴリ，横に 2004 から 2008 年としている．

図にあるように，左列の傾きは強くなるに従って色づけを濃くし，右列の記事の増加率は増加が大きければ大きいほどに色が濃くなるようにした．その結果，傾きに関しては，全体において，年を経るに従って，傾きが強くなるという傾向が明らかとなった．つまり，使用される頻度の高いハイパーリンクはますます高くなっていく，ということである．

4.2 Wikipedia の各カテゴリの成長の傾向

次に，Wikipediaの各カテゴリの個別の成長の傾向について分析する．自己組織化マップによって生成された 9 つのクラスタは，経年のハイパーリンクの分布の係数を変数としてクラスタリングを行っているものである．そのため，5 年間の傾向の似たものが同じクラスタに分類されるようになっており，したがって，このように異なる傾向が現れたということから，成長パターンに多様性がある，と言える．しかし，そこにどのような差異があるために成長パターンに多様性がうまれるのか．それを明らかにするために，新たに 2 つの指標を取り入れた．

(1) 記事増加率

図 1 の各クラスタにある，右列が前年比の記事増加率である．記事の増加が大きければ色が濃くなるよう色づけした．

まず，自己組織化マップの全体像を把握すると，マップの右側に年々傾きが強くなり記事の増加が少なくなる傾向のクラスタ，左側に傾きはまだ大きくなく，記事の増加率は高い水準のクラスタとなっている．また，マップの下側には，カテゴリが 2005 年以降に作成されたクラスタが多く位置づけられている．

近似式の傾きと記事増加率とをあわせてみると，2007～2008 年に傾きが非常に強い C3，C5 のクラスタにおいては，記事の増加率が低いという傾向がある．対して，傾きの小さい C2 は，2008 年にも記事の増加率が大きい．また，マップの下側に位置づけられている C7～9 は，カテゴリが作成された時期が，他のクラスタと比較して遅いため，まだ記事の傾きや増加率の傾向が定まらない不安定な状態と言える．

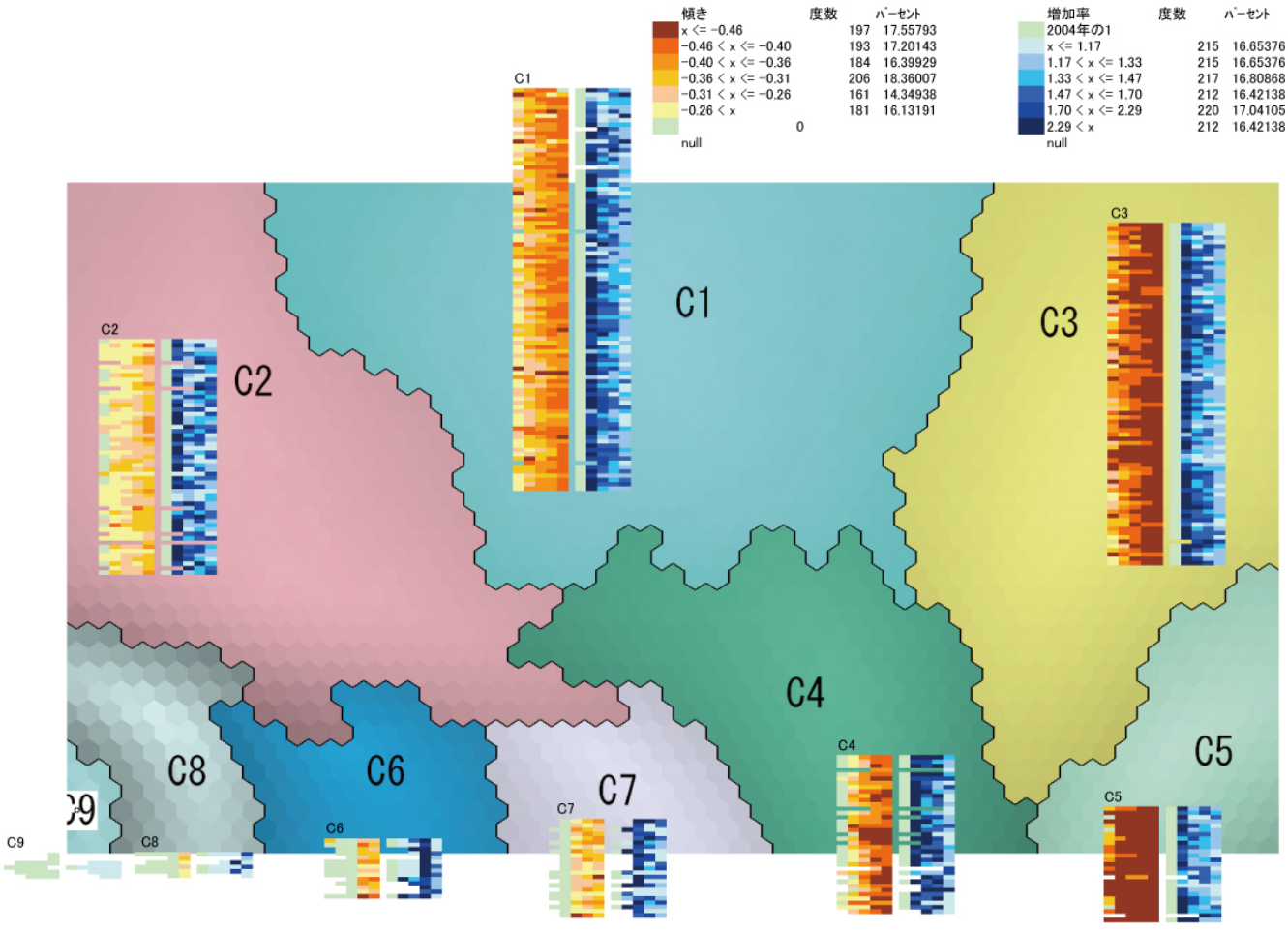

図 1　クラスタリング，および傾き・記事増加率の可視化

クラスタ	科学	学問	技術	自然	社会	人間	総記	地理	文化	歴史	総計
C 1	0.69	0.73	1.10	0.97	1.06	1.29	0.52	0.81	1.62	0.91	1.00
C 2	1.95	0.67	0.94	0.55	0.97	1.47	0.89	0.35	1.31	1.25	1.00
C 3	0.79	1.23	1.03	1.97	1.03	0.99	0.60	0.93	0.20	0.63	1.00
C 4	1.16	1.12	0.70	0.55	1.24	0.55	2.66	1.03	0.86	1.39	1.00
C 5	0.40	2.04	0.96	1.50	0.57	0.75	0.00	2.12	0.00	0.00	1.00
C 6	0.77	0.33	1.85	0.00	0.55	1.45	0.00	2.73	1.71	1.23	1.00
C 7	1.87	1.20	0.56	0.00	0.66	0.00	2.14	1.66	1.04	2.24	1.00
C 8	0.00	0.76	1.08	0.00	2.55	0.00	0.00	0.00	1.33	2.87	1.00
C 9	0.00	0.76	1.08	0.00	1.27	0.00	8.19	0.00	2.67	0.00	1.00
総計	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

1.0より大
1.5より大

表 2　クラスタ×主要カテゴリの特化係数一覧

(2)　特化係数

次に，各主要カテゴリが，どのクラスタに特化しているか，その傾向を見るために特化係数を算出する．表 2 は，クラスタがどの主要カテゴリに特化しているか，の一覧である．

特化係数 S は，以下のように計算した．P はカテゴリ全体を表し、Pm は、主要カテゴリ m を表わす．また，C はクラスタ全体を表し，Cn はクラスタ n を表す．そして，N はカテゴリ数を表し，したがって，$N_{Pm,Cn}$ は主要カテゴリ m かつクラスタ n のカテゴリ数を表す．

$$Smn = \left(N_{Pm,Cn} / N_{Pm,C}\right) / \left(N_{P,Cn} / N_{P,C}\right)$$

これにより，各クラスタがどの主要カテゴリに特化しているかということを見る．特化係数が 1 以上は全体と比較して特化しているということになる．1 より大きい場合は薄いオレンジ，1.5 より大きい場合は濃いオレンジで示している．

先ほど傾きの強かった C3 および C5 は，自然や学問といった体系だったカテゴリに特化していることがわかる．一方，傾きが小さく，記事が依然として増加中の C2 は，科学や人間，文化といったカテゴリに特化している．また，C9 は総記カテゴリに特化している．総記とは，Wikipedia のまとめのような存在のカテゴリであり，増殖する Wikipedia のバインディングのために作られているものである．そのため，他のカテゴリのような秩序が生まれづらく，べき乗にはなりづらいという独特の傾向を持っている．

5.　考察と今後の課題

本分析では，Wikipedia の各カテゴリの記事内で使用されるハイパーリンクの頻度と順位の関係について，経年のデータを使用することで，成長過程に多様性があることを明らかにした．本分析において，明らかになったことを以下にまとめる．

まず，ほぼすべてのカテゴリにおいて共通する特徴として，年を経るごとにべき乗分布の傾きが強くなるという傾向が見て取れた．これは，カテゴリ内において，カテゴリとしての秩序が形成されるためだと考えられる．たとえば複雑系に関するカテゴリは，複雑系カテゴリとして確立されるべきであり，そのため，そのカテゴリに沿ったハイパーリンクをますます使用されるためであると推測される．

次に，様々なカテゴリにおいて，カテゴリの成長過程には差異があるということがわかった．本分析においては，似た成長過程を持つカテゴリ同士をまとめて傾向を把握すべく自己組織化マップによってクラスタリング行った．その結果，経年で傾きが大きくなるカテゴリと，傾きがあまり大きくならないカテゴリがあるということがわかった．さらに，それらを記事の増加率という視点でみると，記事の増加が高い水準で行われている傾向のあるクラスタは，そうでないクラスタと比較して傾きが大きくないままであるということがわかった．これは，記事が高い水準で増加し続けている最中の場合には，カテゴリ内における秩序もまた形成されている途中であり，したがってべき乗分布の傾きは，まだ強くならないためである，と考えられる．また，各クラスタがどの主要カテゴリに特化しているのかをみると，体系だった背景のある学問をはじめとするカテゴリは，傾きが強いクラスタに特化し，反対にそうでない傾向をもつクラスタは，日々新しい項目が増え続ける科学や文化といったカテゴリに特化するということがわかった．これらの記事増加率および特化係数をあわせて考察すると，既に完成された項目に関するカテゴリは，体系がしっかりとしているために傾きが大きく，記事は一度まとめて増加した後には大きく増えることはあまりないということがわかる．反対に，項目それ自体が新規性のあるものであるとき，そこにはまだ確立された体系はないため，ハイパーリンクのヘッドとテールの差がさほど大きくないということがいえる．

本研究では，上記の考察を，いくつかの例によって具体的に把握することを試みた．今後は，成長過程の多様性において，他にどのような変数が有効であるのかということを明らかにし，成長のアルゴリズムに迫るとともに，各クラスタの特徴を，より具体的に理解するための方法を模索することを課題としたい．

参考文献

[井庭 2006]　井庭崇，深見嘉明，斉藤優: 書籍販売市場における隠れた法則性，情報処理学会　第 61 回数理モデル化と問題解決研究会，2006 年 9 月．

[徳高　2007]　徳高平蔵，大北正昭，藤村喜久郎: 自己組織化マップとその応用，シュプリンガー・ジャパン，2007 年 7 月．

[Barabasi　2002]　Barabasi,A.-L.: LINKED: The New Science of Networks, Perseus Book Group, Perseus Book Group, 2002.