社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－MPS－67 (43)
2007－BIO－11 (43)
2007／12／21

Wikipediaの経年変化に関するカテゴリ間の比較分析

山崎由佳† 伊藤貴一†
井庭崇†† 熊坂賢次†††

本論文の目的は，Wikipediaの知の増殖の法則を探ることにある．本論文では，Wikipediaの増殖の法則を明らかにするため，ページタイトル及びページ内のリンクをデータとして用い，時系列の分析を行う．その際，Wikipedia内でのカテゴリ間の差異を検証するため，特徴的な4カテゴリを対象とし，それぞれの経年データによってカテゴリ間の比較分析をする．その結果，ページ自体の増殖においては新しい事象に関するカテゴリが高い水準という一般的な結果であったが，一方，ページ内のリンクの増殖という，ページの豊かになっていく様については，カテゴリの内容が集合知的か否かによって特徴的な法則があることがわかった．

Comparative Analysis of four Categories' Growth in Wikipedia

Yuka Yamazaki,† Takaichi Ito,† Takashi Iba††
and Kenji Kumasaka†††

In this paper, we explore the rule of the growth of Wikipedia. To make clear the growth rule of Wikipedia, we use pages' title and links as data, and analyze growth. In this occasion, we choose four characteristic categories to analyze difference during them. In the result, analysis of pages' growth lead new content categories growth in high level, in the other hand, we found that contents of categories much influence links' growth.

1. はじめに

近年，WWW上で，見知らぬ人同士のコラボレーションによってコンテンツ作成が行われており，その代表例にWikipedia (http://ja.wikipedia.org/) がある．そこでは，無限のスペースの中で自己増殖的に知が編纂されている．本論文では，増殖の仕方がコンテンツの内容によってどのように異なるのかを分析する．

Wikipediaでは，辞典のひとつひとつの項目がページとして存在している．各ページでは，カテゴリでまとめられ，紙上の辞典同様に文や写真などによって項目が説明されている．文中の特徴語は任意でハイパーリンクとすることが可能であり，それをクリックすると，ハイパーリンク語の項目ページへとジャンプすることが出来る．

これまで，ブログ記事ネットワークからの特徴的トピックの時系列分析[1]や，時系列テキスト集合からの社会的関心の分析[2]など，ウェブ上のデータの時系列分析が行われてきた．本研究では，Wikipediaにおいてカテゴリごとの比較分析を行うことで，記事コンテンツの内容が時系列の変化にどのような影響を及ぼすかを分析する．

† 慶應義塾大学 政策・メディア研究科
Graduate School of Media and Governance, Keio Univ.
†† 慶應義塾大学 総合政策学部
Faculty of Policy Management, Keio Univ.
††† 慶應義塾大学 環境情報学部
Faculty of Environment and Information Studies, Keio Univ.

2. 分析の対象と方法

本研究では，Wikipediaのページをノード，ページ内のハイパーリンク（以下リンク）をページの保有する変数と見なし，データを収集した．対象となるデータは，2004年より各年の1月1日時点で表示されるページを履歴より取得した．なお，データ収集は2007年9月に行った．

ページ間の共起の成長を概観するために，ページをノード，ページをつなぐハイパーリンクを変数とし，各ページ間で共通リンクが1つ以上ある場合にページ間を線でつなぐ2部グラフのネットワークを作成し，その解析を行う．

また，ページ間の関係を具体的なページ名として可視化し，解釈するために，自己組織化マップ (SOM: Self-Organizing Maps) を用いて分析する．SOMと

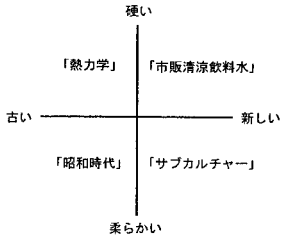

図 1　分類軸と対象カテゴリ

は，競合学習を基礎とした人工ニューラルネットワークの一種であり[3]，関係の深いノード同士がひきつけ合う．このアルゴリズムを利用し，ハイパーリンクを変数としてページを球面にプロットし，全体像を可視化する．

3. 各カテゴリの分析

本研究では，**図 1** の 4 カテゴリを分析対象とする．図 1 において縦軸の「硬い－柔らかい」という指標は「硬い」側には科学的裏づけの強いコンテンツを置き，対して「柔らかい」側に多様な解釈可能なコンテンツを想定した軸である．これは，Wikipedia が人々の集合知として認識・機能しているという意味で，硬い（辞書的）か柔らかい（集合知的）か，という重要な指標であると考えた．横軸の「古い－新しい」という指標は，コンテンツそれ自体の発生した時間である．

3.1 熱力学カテゴリ

ページ数の経年変化に関しては，当初は増加が著しいものの，年を経るごとに水準が低くなる（**表 1**）．次に，ネットワーク構造の推移に関しては，2004 年時点で 0.252 だったクラスタリング係数は，0.371，0.435，0.465 と徐々に上昇し，近隣ページ数平均も同様に 7.852，27.24，56.896，78.732 と上昇していく．経年で徐々にネットワークが成熟していくことから，本カテゴリでは，同リンク参照が経年で強まるという傾向があると言える．

球面 SOM の分析においては，**図 2** より，一度近隣にプロットされたページ同士は，以降も近隣に関係するということがわかる．これは，熱力学カテゴリのページが，経年で大きく変化することの少ない内容であるためだと考えられる．また，「法則」や「サイクル」などの共通の項目は，本カテゴリにおいてはある程度ページの内容も共通性があるということが見て取れる．

3.2 昭和時代カテゴリ

ページ数の経年変化に関しては，熱力学カテゴリ同様，当初は増加が著しいものの，徐々に増加率は低く

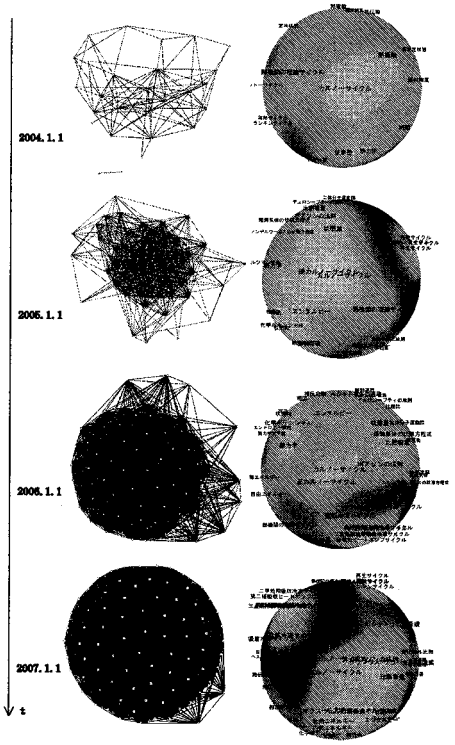

図 2　熱力学カテゴリのネットワーク図（左）および球面自己組織化マップ（右）

なる．次に，リンク数の経年変化については，年を経るごとに平均リンク数は増加し，標準偏差も大きくなる．これは，多様な解釈および厚みのある記述を昭和時代カテゴリというコンテンツが求める所以である．

共起ネットワークにおいては，2004 年より 0.335，0.341，0.390，0.425 と変化し，近隣ページ数平均は 5.6，17.28，39.377，60.660 と変化する．**表 2** より，本カテゴリはリンク数の増加が著しく，それに伴ってページ間の関係が密になることが伺える．

球面 SOM の分析においては，その年毎に近隣ページが大きく異なる点である．これは，多くの人々によって多様に執筆され続けるためであると考えられる．

表 1　熱力学カテゴリのページ数およびリンク数の経年変化

日付	ページ数	ページ数 前年比	リンク数 平均	リンク数 標準偏差
2004	29		9.59	5.94
2005	53	182.8 %	12.38	9.04
2006	68	128.3 %	16.91	14.42
2007	83	122.1 %	19.17	10.83
平均	58.25	144.4 %	14.51	10.83

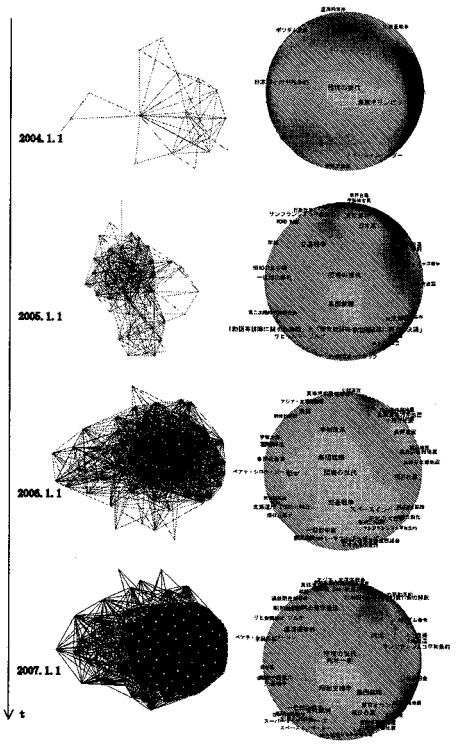

図 3　昭和時代カテゴリのネットワーク図（左）および球面自己組織化マップ（右）

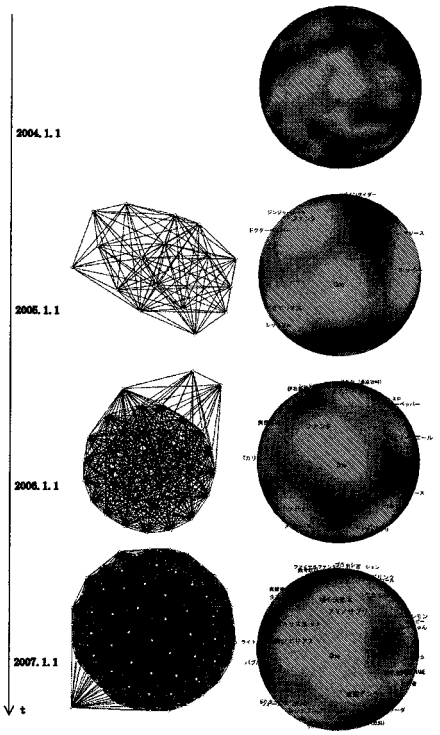

図 4　市販清涼飲料水カテゴリのネットワーク図（左）および球面自己組織化マップ（右）

3.3　市販清涼飲料水カテゴリ

ページ数については，経年で高い水準で増加していることが特徴である．これは，日々対象となる項目自体が生まれ続けるカテゴリであるためと考えられる．対して，リンク数の経年変化は比較的小さい（**表 3**）．

次に，共起ネットワークにおけるクラスタリング係数は 2005 年より 0.446，0.480，0.496 とほぼ変わらず，近隣ページ数平均は 2005 年より 12.941，29.455，59.935 となっている．本カテゴリにおける特徴は，クラスタリング係数が経年でほぼ変わらないという点である．これは，ページが作られた当初から豊富なデータがあり，かつそれがメーカー名など広告としての相互参照が多いためであり，加えて，それ以降の変化が少ないためであると考えられる．

また，球面 SOM においては，熱力学カテゴリ同様，一旦両者の位置関係が決まると，以降の関係は大きく変わることはない．

3.4　サブカルチャーカテゴリ

ページ数については，市販清涼水カテゴリ同様，経年で高い水準で増加し続けている．また，リンク数も経年で大きく増加する（**表 4**）．本カテゴリで最も特徴的なことは，共起ネットワークにおいて 2004 年時点でクラスタリング係数が 0.0 ということである．そして経年で，2004 年より 0.0，0.221，0.334，0.462 と

表 2　昭和時代カテゴリのページ数およびリンク数の経年変化

日付	ページ数	ページ数 前年比	リンク数 平均	リンク数 標準偏差
2004	21		16.24	23.40
2005	55	261.9 %	33.89	45.48
2006	79	143.6 %	54.20	74.85
2007	99	125.3 %	61.62	80.19
平均	63.5	177.0 %	41.49	55.98

表 3　市販清涼飲料水カテゴリのページ数およびリンク数の経年変化

日付	ページ数	ページ数 前年比	リンク数 平均	リンク数 標準偏差
2004				
2005	19		19.58	13.15
2006	34	178.9 %	23.38	18.26
2007	63	185.3 %	29.70	25.80
平均	38.6	182.1 %	24.22	19.07

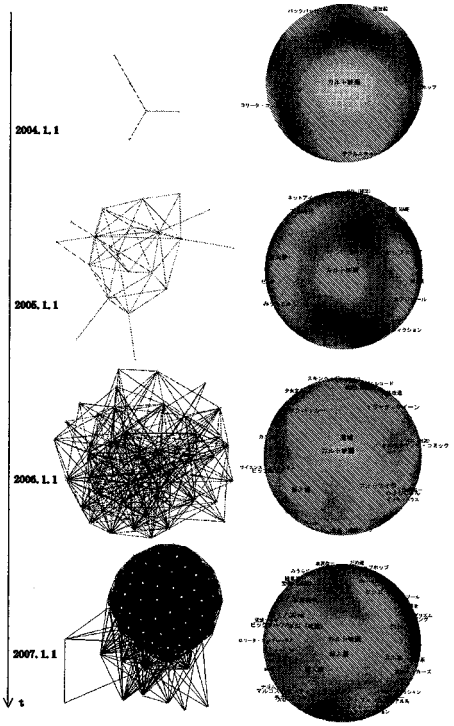

図 5　サブカルチャーカテゴリのネットワーク図（左）および球面自己組織化マップ（右）

上昇している．また，近隣ページ数平均は 1.6，5.3，17.045，56.603 となっている．これは，2004 年時点では，ページ間の共通性がほとんどない状態だったが，年を経るにつれてページが増殖し，増殖するに従って相互参照がなされたためにこのように経年で変化をしていったと考えられる．

また，球面 SOM では，昭和時代カテゴリ同様経年でのページ間の関係が大きく変化する傾向がある．

4. 考　　察

このように，ページ数の増加に関してはカテゴリにおいてページとなる事象自体の新しさが重要な要素となることがわかった．これは，項目自体が日々増えるためであると考えられる．また，リンク数の増加に関しては，昭和時代およびサブカルチャーカテゴリのように，集合知的な柔らかい内容をコンテンツとするカテゴリは，辞書的な硬い内容をコンテンツとする熱力学および市販清涼飲料水カテゴリよりも水準が高いことが明らかになった．

表 4　サブカルチャーカテゴリのページ数およびリンク数の経年変化

日付	ページ数	ページ数 前年比	リンク数 平均	リンク数 標準偏差
2004	10		16.10	16.62
2005	24	240.0 %	31.88	49.83
2006	47	195.8 %	46.30	51.52
2007	74	157.4 %	66.74	68.41
平均	38.75	197.8 %	40.25	46.60

そして，ネットワーク構造の経年変化という観点から考えると，古い内容をコンテンツとする熱力学カテゴリおよび昭和時代カテゴリは，球面自己組織化マップを用いた内容に踏み込んだ解釈においては差異があるものの，ネットワークの経年変化においては似た変化をたどることがわかった．一方，コンテンツの新しいそれぞれ 2 カテゴリについては，内容の硬い市販清涼飲料水カテゴリは作成当初からの豊かなデータのために，すでにページ間の関係は構築された状態となっているが，その後の成長はあまりないという傾向がある．対して，内容の柔らかいサブカルチャーカテゴリは，ページが作成された当初はページ間の関係性が薄いが，経年で豊かになり続ける記述によってページ間の関係が日々強まり続けるということが明らかとなった．

5. おわりに

本研究の今後の発展については，より大量データを対象とした分析による法則性の発見および強化を目的としている．今回，カテゴリにおいて項目が集合知的か否か，また項目自体が日々生まれ続けるか否かという視点から，特徴的な 4 カテゴリの分析を行ったが，他の多くのカテゴリの詳細な分析を以て，Wikipedia の経年変化の傾向における法則性を見出したい．

参 考 文 献

1) 内田誠, 柴田尚樹: ブログ記事ネットワークからの emerging topic の抽出と可視化, 人工知能学会第 20 回全国大会 (2006).
2) 福原知弘, 中川裕志, 西田豊明: 時系列テキスト集合からの社会的関心の分析, 日本機会学会第 16 回インテリジェント・システム・シンポジウム (2006).
3) 徳高平蔵, 大北正昭, 藤村喜久郎 (編): 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).