

F-005

サーバ所在区分別で見た Web グラフ構造の分析 ～outdegree 成長図の利用～

An analysis of Web graph by server location category.

難波 弘行[†]

Hiroyuki Namba

中平 勝子[‡]

Katsuko T. Nakahira

三上 喜貴[‡]

Yoshiki Mikami

1 はじめに

インターネットは、今や世界的に利用される一般的な通信手段となったが、国によって通信インフラの整備状況やインターネット上の検閲・規制が影響し、インターネットの発達に大きな相違が生じていると考えられる。

インターネットの発達の度合を観測する手段のひとつとして、インターネット上のハイパーリンクのグラフ接続に注目した outdegree 分布図を用いる方法がある。石原[1]は国を示す単位として ccTLD を採用し、ccTLD 毎に outdegree 分布図による分析を行い、Web 空間上での活動に対して規制が少なく自由な成長が許容されている国には outdegree 分布図にべき乗則が現れるということを示している。

しかし、URL に書かれた ccTLD とサーバの設置場所は必ずしも一致しないため、ある国の ccTLD において Web ページやリンクの総数が多かったとしても、それが純粋な国内での発達かどうかという点で疑問があった。また、国内外でのサーバの物理的所在割合に対する統計はなされているが[2,3]、outdegree 分布図による分析はなされていない。そこで、本稿では ccTLD に加えサーバ所在地という視点を加え、サーバ所在地によって国内・国外・不明の3区分別で outdegree 分布図を作成し、各国の Web 利用形態の相違を明らかにすることを試みた。

2 方法

本研究では、ミラノ大学の開発したクローリングロボットである UbiCrawler[4]を用いて 2006 年の 7 月に収集したアジア地域 31 ヶ国（日本、韓国、中国を除く）のデータを使用した。「.com」や「.net」のようなジェネリック・トップレベルドメインは、その帰属する国の特定が困難であるため、国の特定は ccTLD を単位として行った。

また、サーバ所在地情報の取得には Maxmind 社が提供している IP アドレスから地理情報を得る仕組みである GeoIP[5]を用い、各 ccTLD のページをそれらのサーバ所在地から国内・国外・不明の3つの区分に分割した。

指標として用いた outdegree とは、グラフ理論における用語であり、有効グラフにおいて1つの頂点から出て行く枝の総数を示している。Web 空間を巨大な有効グラフとしてとらえたとき、outdegree はある Web ページから他の Web ページに対して張られているリンクの総数を表す。本稿で用いる outdegree 分布図は、横軸に outdegree 値、縦軸にその outdegree 値を持つページの総数をとり、両対数グラフで示したものである。Broder ら[6]は、この outdegree 分布図の形状はほぼべき乗則に従うことを示している。べき乗則は基本的に構造安定性を示すものとされている。

また、石原[1]は outdegree 分布図を国の総ページ数の尺度に沿って整理したとき、outdegree 分布図は、Web ページの総数が大きくなるにつれ、べき乗則に従う分布へと

近づき傾向があると示している。これは、共時的な分析ではあるが、一国について通時的に追跡してその成長を確認するためには長い時間を要してしまう。そこで、本研究では outdegree 分布図を Web ページの総数順に整理し、通時的な Web 空間の発達過程の一部としてとらえたものを outdegree 成長図と定義する。

3 結果・考察

3.1 outdegree 分布図による分析

各 ccTLD について、サーバ所在区分が国内、国外、不明の3種類の outdegree 図を作成したところ、いくつかの特徴的な傾向が見られた。以下では、それについて outdegree 分布図の例を用いながら説明する。

1)国内集中型

図1の左端に示したパプアニューギニアでは、国外や所在地不明サーバがいくらか存在するものの、全体としては国内サーバがその大半を占めている。この国内集中型は、Web ページの総数が比較的少ない国においていくつか見られ、このような国では Web 空間は国内サーバを中心に成長していると考えられることができる。また、パプアニューギニアの国内サーバのグラフに垂直線状の形が見受けられるが、これは掲示板のように動的に生成される Web ページが示す特徴的なパターンであり、同じようなリンク数を持つページが大量に存在する場合に現れる。パプアニューギニアの場合、全体の 65%が政府機関に属する Web ページであり、それらは CMS を用いて大量に動的生成されていると考えられる。

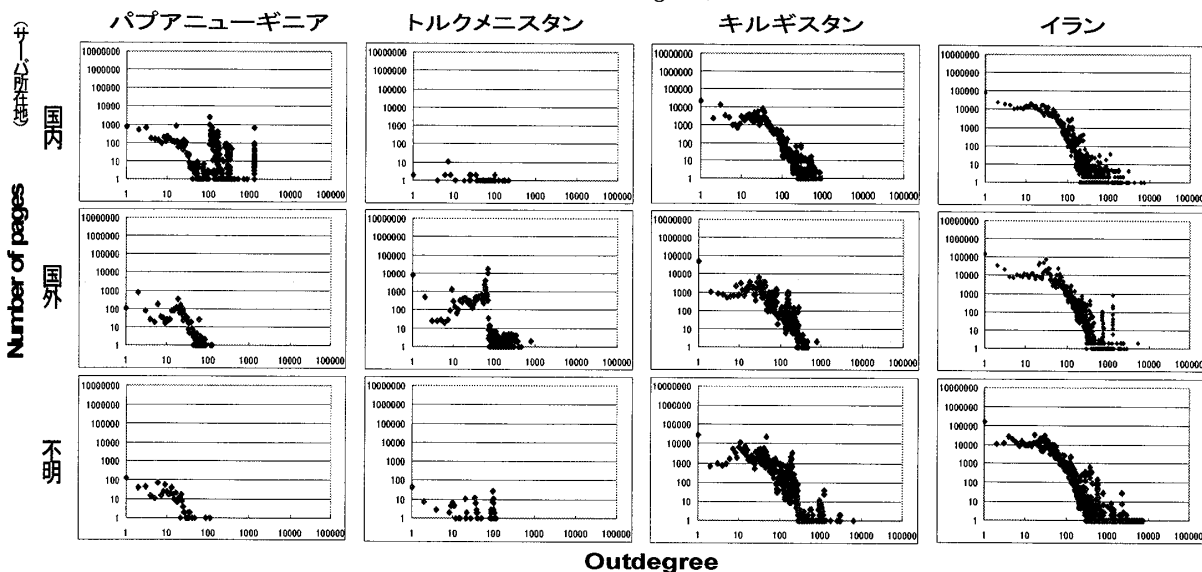
2)他国依存型

図1の左から2列目に示したトルクメニスタンでは、国内にはほとんどサーバが無く、国外のサーバが多く存在している。このようなグラフが見られる国は、国内にサーバを設置するのが難しい国や、ドメインの使用権を海外に販売している国であることが考えられる。トルクメニスタンの場合、「.tm」が過去に商標 (trademarks) ビジネスのために用いられたという経緯があるため、国外のサーバが多く見られる結果になったと考えられる。

3)均等発達型

図1の左から3列目に示したキルギスタンでは、国内、国外、不明サーバが共にバランス良く存在し、それぞれのグラフは outdegree 値が 30 程度までは目立った推移が無く、それを越えるとべき乗則に従うような直線を形成しているのが伺える。このような均等に発達したグラフは、Web ページの総数が上位の国ほどよく見られ、これらの国は国内外共に自由な成長を遂げていると考えられる。

図1.outdegree 分布図



4)成熟型

図1の右端に示したイランは、どの所在区分に関しても大きな差が無い点に関しては前述の均等発達型とほぼ変わらないが、outdegree分布のばらつきが小さくなっており、おおよそどの所在区分のグラフについてもある一定の形に近づいているように見受けられる。また、このような成熟型の国のoutdegree分布図には、動的生成ページと思われる垂直線が国内外で多く現れる傾向があった。これは、Web空間の発達が進むとポータルサイトのような比較的規模の大きいWebサイトが出現しやすくなることによるのではないかと考えられる。

3.2 所在区分別のべき乗則評価

表1は、各国の自国サーバと他国サーバのoutdegree分布図に対し、それぞれべき乗則が確認できないものをI、べき乗則とみられる分布を形成しているものをIIとして評価を行い、分類したものである。

また、今回はoutdegree値が30程度までの分布にばらつきがあっても、それ以降でべき乗則らしき規則性が見られるものはIIとして評価した。

表1.べき乗則評価結果

国外 / 国内	I	II
I	ミャンマー, トルクメニスタン	東ティモール, アフガニスタン, ラオス
II	パプアニューギニア, イエメン, モルディブ, ブータン, シリア, カタール, パレスチナ, ブルネイ, スリランカ, オマーン	クウェート, カンボジア, バングラデシュ, タジキスタン, レバノン, ヨルダン, バーレーン, レバノン, ネパール, モンゴル, キプロス, パキスタン, キルギスタン, アラブ首長国連邦, サウジアラビア, アゼルバイジャン, イラン

表1を見ると、Webページ総数下位の国（下から東ティモール、ミャンマー、パプアニューギニアなど）は国内あるいは国外でべき乗則が見られず、国内集中型や国外依存型と見られる国が多い。一方、Webページ総数上位の国（上からイラン、サウジアラビア、アゼルバイジャンなど）は国内外ともにべき乗則らしき分布が見られたため、II-IIに多く分類される結果となった。ここから、Web空間のある程度の発達には国内サーバと国外サーバの両方の存在が関わっていると言えるだろう。

以上を踏まえ、Webグラフ構造は各国の事情にあわせて国内もしくは国外サーバにてまず発達し、インターネットの普及・大衆化につれ国内外ともに構造安定化していくと考えられる。

4 まとめ

本稿では、インターネット上に存在するWebページをccTLDに加えてサーバの所在区分で分けてoutdegree分布図を作成し比較を行うことで、それらが各国のインターネットの利用形態やWeb空間の成長の過程を知るための手がかりと成り得ることを示した。今後は、いくつかのoutdegree分布図上に現れたべき乗則と乖離した分布や、所在地不明サーバについての調査を進めていく予定である。

参考文献

[1]石原直幸:グラフ理論を用いたカントリードメインのリンク構造解析,長岡技術科学大学大学院 工学研究科 修士課程修士論文, 2008
 [2]Nakahira,K, Hoshino,T, And Mikami,Y, "Geographic Location of Web Servers under African Domains", In Proc. of WWW2006, pp.989-990(2006)
 [3]平手勇宇,片瀬弘晶,山名早人:"全世界のWebサーバの地理的位置・バックリンク数の解析",日本データベース学会論文誌,Vol.7,No.2,pp.1-6(2007)
 [4]Ubicrawler <http://law.dsi.unimi.it/>
 [5]GeoIP <http://www.maxmind.com/>
 [6]A.Broder, R.Kumar, F.Maghouli et. al. : Graph structure in the web, The 9th International World Wide Web Conference, May 2000

†長岡技術科学大学 Nagaoka University of Technology