

# 全世界の Web サーバの 地理的位置・バックリンク数の解析

平手 勇宇<sup>†</sup> 片瀬 弘晶<sup>††</sup> 山名 早人<sup>†††</sup>

Web サーバからは、膨大な情報が発信され続けており、我々の調査では 2005 年 10 月時点で世界中に 537 億のページが存在すると予測している。我々はこれまでに e-Society プロジェクトで収集した約 107 億ページのトップレベルドメイン分布、言語分布について調査を行ってきた。本稿では、これまでの解析の発展として、Web サーバの地理的設置位置分布、Virtual Host 数の地理的分布、および Web サーバ単位でのバックリンク数分布の解析を行った。その結果、約 95.5% の Web サーバが、北米・ヨーロッパ・アジア地域に設置されていること、1 ホストあたりの Virtual Host 数が多い国は、中南米・東欧諸国であること、そしてバックリンク数と Web サーバ数は Power-Low に従っていることを確認した。

## Geographical Location and Number of Back-Links of Web Servers All Over the World

Yu HIRATE<sup>†</sup> Hiroaki KATASE<sup>††</sup> Hayato YAMANA<sup>†††</sup>

According to our investigation result in Oct. 2005, the number of Web pages all over the world is estimated 53.7 billion. We have investigated TLD distribution and Language Distribution of Web pages based on 10.7 billion Web page dataset. In this paper, as one of our Web statics investigation series, we analyzed three kinds of distribution based on 10.7 billion Web page dataset, distribution of geographical location of Web server, the number of virtual hosts per one Web server, and the number of back links, i.e. the value of in-degree, per one Web server. Our results show (1) about 95.5% of Web servers are located in North America, Europe, and Asia regions, (2) hosts located in Latin America and East Europe have a large number of virtual hosts, and (3) the distribution between the value of in-degree and the number of Web servers follow the power law.

### 1. はじめに

近年、Web サーバから発信される情報量が膨大になり、2005 年 10 月時点で世界中に 537 億のページが存在するという調査結果が報告されている<sup>1)</sup>。このような大量の Web ページから有用な知識を抽出することを目的とし、リンク解析をはじめとする Web マイニングに関する研究<sup>2)3)</sup>が幅広く行われている。しかし、統計的に意味のある解析を行うためには、Web についての各種統計的な分布を知ることが欠かせない。

これまでに、Web の各種統計に関する様々な調査・研究がおこなわれている。2001 年に池内らは、検索エンジンを用いて、トップレベルドメイン(=TLD)分布状況、および Web ページの言語分布状況を統計的に調査している<sup>4)</sup>。2006 年に Bar-Yossef らは、検索エンジンのサンプリング手法を利用して、Web ページの TLD 分布について言及をしている<sup>5)</sup>。しかし、検索エンジンを利用していることから、予測された分布状況は検索エンジンのインデックスの偏りに依存するという欠

点をもつ。

実際の Web データを対象とした解析として、著者らのメンバーである加藤らは、2006 年に e-Society プロジェクト<sup>6)</sup>で収集を行った com と jp ドメインを主対象とした約 30 億ページデータをもとに、TLD 分布、言語分布、およびリンク解析を実施した。しかし、解析対象の Web ページは、com、jp に偏っており、Web 全体を網羅した解析にはなっていない。

そこで我々は、2007 年度より、同じく e-Society プロジェクト<sup>6)</sup>で収集した約 107 億ページを対象として解析を開始した。これまでに、Web ページの TLD 分布、言語分布についての解析を行った<sup>8)9)</sup>。その結果、「com」に属する Web サーバが全 Web サーバの 37.8%、英語で記述されたページが全 Web ページの 42.6%と、現在の Web が「com」、「英語」に偏っていることを示してきた。

本稿では、8)9)の解析の発展として、約 107 億ページが配信されている Web サーバの地理的設置位置分布、および Web サーバ単位のバックリンク数(=In-degree 数)の分布についての解析結果を報告する。本稿では、以下のような構成を取る。第 2 節において、Web サーバの地理的設置位置に関する関連研究、および Web ページの In-degree 数分布に関する研究を示す。第 3 節において、解析対象のデータセットと解析手法について示す。第 4 節において、解析結果を示し、第 5 節にてまとめを行う。

<sup>†</sup> 早稲田大学メディアワークセンター  
Media Network Center, Waseda University

<sup>††</sup> 早稲田大学基幹理工学研究所  
Graduate School of Fundamental Science and Engineering, Waseda University

<sup>†††</sup> 早稲田大学理工学術院・国立情報学研究所  
Science and Engineering, Waseda University /National Institute of Informatics

## 2. 関連研究

### 2.1 Web サーバ地理的位置特定に関する研究

中平らは、2006 年にアフリカの国別ドメイン (=ccTLD)に属する 1,600 の Web サーバに対し、当該 Web サーバの設置位置の特定を行っている<sup>10)</sup>。10)では、自国内に設置されていたホストは約20%であり、残りの約80%は主に欧州、北米に設置されていたという報告がされている。また、2007 年に近藤らが IP アドレス 100 万件を対象に、Web サーバの設置位置を特定した結果、約52.8%の Web サーバが米国内に設置されていたという結果を報告している。

近藤らの研究<sup>11)</sup>では、URL を IP アドレスに変換し、Web ページが属する Web サーバの地理的な位置を特定している。11)では、任意のページにリンクしている Web ページ集合の地理的位置を特定し、地理的位置の分散状況によって当該ページの地域性を評価している。近藤らは、IP アドレスをランダムに 100 万件を生成して調べた結果、アメリカが1位で58.17%、日本が2位で6.06%である結果を報告している。

このように、Web サーバの設置位置を特定する研究は行われているが、調査対象の Web サーバ数が少なかったり、地域的に偏っている問題点をもつ。これに対して本稿では、e-Society プロジェクト<sup>9)</sup>で収集した約 107 億ページの Web データを用いて、全世界の Web サーバの地理的設置位置の特定を行う。

### 2.2 Web サーバ地理的位置特定に関する研究

これまで、Web ページのバックリンク数 (=In-degree 値)の分布に関する様々な研究が行われている。すべての研究において、Web ページの In-degree 値と、Web ページ数の分布は Power-law に従っているとの結果が報告されている。一般に、Power-law に従っている場合、In-degree 値  $x$  と In-degree 値に対応する Web ページ数  $y$  の関係において、式(1)のような関係が成立する。

$$y = ax^{-\gamma} \quad (1)$$

ここで、 $a$ は定数を表し、累乗数 $\gamma$ は Power-Low の強さを表す。

Kumar らは 1999 年に Alexa で収集した約 20 万ページの Web ページデータを対象に $\gamma$ の値を計算した結果、おおよそ 2 であると結論づけている<sup>12)</sup>。また、Barabashi らは同年に、約 80 万ページ $a$ に対して計測した結果、 $\gamma=2.1$ であったと報告している<sup>13)</sup>。Border らは 2000 年に 14)において、AltaVista の約 20.3 万ページのデータに対して適用させた結果、 $\gamma = 2.1$ であったと報告している。Boldi ら<sup>15)</sup>は、2002 年にアフリカの約 200 万ページに対して解析したところ、 $\gamma = 1.92$ 、Liu らは 2005 年に、Peking University Sky net Search Engine で収集した約 140 万ページを基に、 $\gamma = 2.05$ であった<sup>16)</sup>とそれぞれ報告している。以上をまとめると、表 1 に示す通りとなる。

これらの既存研究は、対象とする Web の年代が古い

表 1 In-degree に関する既存研究

著者	年	ノード数	$\gamma$ の値
Kumar ら <sup>12)</sup>	1999	Alexa で収集した 20 万ページ	2 (roughly)
Barabasi ら <sup>13)</sup>	1999	80 万ページ	2.1
Broder ら <sup>14)</sup>	2000	Altavista で収集した 20.3 万ページ	2.1
Boldi ら <sup>15)</sup>	2002	アフリカの Web ページ 200 万ページ	1.92
Liu ら <sup>16)</sup>	2005	Peking Univ. Sky Net Search Engine で収集した 140 万ページ	2.05

という問題点や、対象 Web が地域的に偏っているという問題点が存在する。これに対して、本稿では 2006 年 7 月までに収集した全世界の 107 億ページデータを対象に、 $\gamma$ の値を示す。

## 3. 解析手法

### 3.1 データセット

解析対象の Web ページは、e-Society プロジェクト<sup>9)</sup>によって収集した Web ページである。e-Society プロジェクトでは、2004 年 1 月から Web ページの収集を継続しており、2004 年 1 月～2006 年 7 月に合計で 14,456,201,906 ページの収集を行った。Web ページの収集にあたっては、2004 年 1 月の収集開始時に、著者らが保有していた com, org, edu, net, uk, jp, us, ca, at の 9 つの TLD から合計 600 万の Web サーバリストを起点として設定した。起点からは最大 15 ホップ先までを収集し、新規に発見した Web サーバは起点として随時追加を行った。なお、Web サーバに与える負荷を軽減するため、1Web サーバあたりの最大収集ページ数を 5 万ページとすると共に、CGI ページを発見した場合は、当該 CGI ページからのホップ数を 1 に制限した。

2006 年 7 月時点において、世界中で発見した Web サーバ数は約 13,468 万台であり、内 8,116 万台の収集を完了した。なお、8,116 万台の内、実際に収集できたのは 5,548 万台であり、2,568 万台 (収集済サーバの約 32%) は既に IP アドレスが存在しない等の理由からアクセスができなかった。また、256 万台 (収集済 Web サーバの 3%) については、/robots.txt によりクローラによるアクセスが禁止されていた。残り約 5,000 万台については未収集である。

2006 年 7 月時点における Netcraft 社<sup>b)</sup>が発表している統計情報によれば、全世界の Web サーバ数 $c$ は約 8,761 万台であり、我々のクローラは、Netcraft 社が把握している Web サーバ数の約 1.5 倍の Web サーバを発見する

b <http://www.netcraft.co.uk/>

c Netcraft 社では、同一 IP でも異なるドメイン名を持つ場合、ドメイン名の異なり数を Web サーバ台数としている。また、既にアクセスできなかったドメイン名も全 Web サーバ台数に加えている。この算出方式は、本論文での算出方式と同一である。

a 対象データの詳細については記述されていない。

ことに成功している。

以下では、2006年7月までに収集された約144.5億Webページの中から10,696,996,553ページdを対象として解析を行った。

### 3.2 Webサーバの設置位置の特定

全世界を、経度2度、緯度2度四方セルに分割を行い、あるWebサーバがどのセルに設置されているのかの特定を行った。その後、全世界のWebサーバ設置位置、およびTLD毎のWebサーバ設置位置の傾向を把握するために、全世界のWebサーバ集合、TLD毎のWebサーバ集合の母集団を定義し、任意のセルに設置されているWebサーバ数のカウントを行った。

### 3.3 バックリンク数の計算

約107億ページのバックリンクの算出を短時間で行うことは困難である。そこで、本稿ではWebサーバ単位でのバックリンク数の算出を行った。つまり、リンク先のWebサーバが、リンク元のWebサーバと違ったリンクのみを、バックリンク数計算対象のリンクとした。

## 4. 解析結果

### 4.1 Webサーバの設置位置分布

#### 4.1.1 全世界のWebサーバの設置位置分布

解析対象の約4,700万台のWebサーバ全てに対し、IP-経度・緯度変換テーブル<sup>17)</sup>を利用して、Webサーバの設置位置の分布を求めた結果を図1に示す。図1右上の「Total」の数字は、我々が収集を行った全てのサーバ数を示しており、「Unknown」の数字は、設置場所を特定できなかったWebサーバ数を示す。なお、設置場所が特定できなかった理由は、ドメインからIPアドレス解決できなかった、もしくはIP-経度・緯度変換テーブルでのエントリがなかった理由が挙げられる。

図1に示す通り、Webサーバの大部分が、米国・欧州・アジアに設置されていることがわかる。表2では、全世界を「アジア」「アフリカ」「オセアニア」「中東」「中南米」「北米」「欧州」「その他」の地域に分類した際のWebサーバ数の分布を示したものである。表2によると、全Webサーバの約50.9%が北米に、約32.1%がヨーロッパに、約12.5%がアジアに設置されており、この3つの地域だけで約95.5%という高い割合を占める結果となった。

表3は、Webサーバ設置数が多かった上位20カ国のWebサーバ数、及び設置数割合を示したものである。Webサーバ設置数が多かった北米地域の米国とカナダは、それぞれ1位と9位であった。Webサーバ設置数が多い上位20カ国は、北米、欧州、アジアに位置する国々で構成されているが、例外はメキシコである。メ

キシコは中南米の国でありながら13位という高い順位となっている。

#### 4.1.2 TLD毎のWebサーバの設置分布

次に我々は、国別トップレベルドメイン(=ccTLD)毎に、当該ccTLDに属するWebサーバの設置位置の分布を計測した。図2にjpドメインに属するWebサーバの設置位置分布を示す。

図2において、「Inside」の数字は、当該ドメインを管理している国内に設置されているWebサーバ数であり、「outside」の数字は、国外に設置されているWebサーバ数である。図2に示す通り、jpドメインに属するWebサーバのうち、約94.1%が自国内に設置されていることがわかる。

ここで、すべてのccTLDに関して、当該ccTLDを管理している国の国内に設置されているWebサーバ数(以下、自国内Webサーバ数とする)、国外に設置されているWebサーバ数(以下、国外Webサーバ数とする。)を計算し、同一グラフにプロットを行うと、図3のグラフが生成できる。

表2 Webサーバ設置数の地域別分布

設置場所	Webサーバ数	割合
アジア	3,962,280	12.51%
アフリカ	48,322	0.15%
オセアニア	223,279	0.71%
中東	106,912	0.34%
中南米	1,046,527	3.30%
北米	16,107,027	50.87%
欧州	10,169,807	32.12%
その他	1,187	0.00%

表3 Webサーバの設置数が多い上位20カ国

設置場所	Webサーバ数	割合
United States	15,325,293	48.8%
Japan	2,437,179	7.69%
United Kingdom	2,060,505	6.51%
Germany	1,937,323	6.12%
Poland	1,048,544	3.31%
Russian	889,628	2.81%
China	830,577	2.62%
Netherland	827,878	2.61%
Canada	780,223	2.46%
France	730,039	2.30%
Brazil	500,162	1.58%
Italy	469,912	1.48%
Mexico	455,267	1.44%
Denmark	335,511	1.06%
Switzerland	235,851	0.74%
South Korea	200,617	0.70%
Taiwan	217,768	0.69%
Sweden	216,908	0.69%
Czech Republic	212,351	0.67%
Spain	198,380	0.63%

d Disk故障等により一部のデータが利用できなくなり、解析にあたっては107億ページを利用した。

e 経度2度緯度2度四方の領域は、おおそ関東甲信越地方の大きさと同等である。

f 「その他」の地域とは、大西洋・インド洋・太平洋の中央に位置する島国、および南極が含まれる。

g すべてのTLDのWebサーバ設置位置分布は、<http://www.yama.info.waseda.ac.jp/e-society/>にて公開中である。

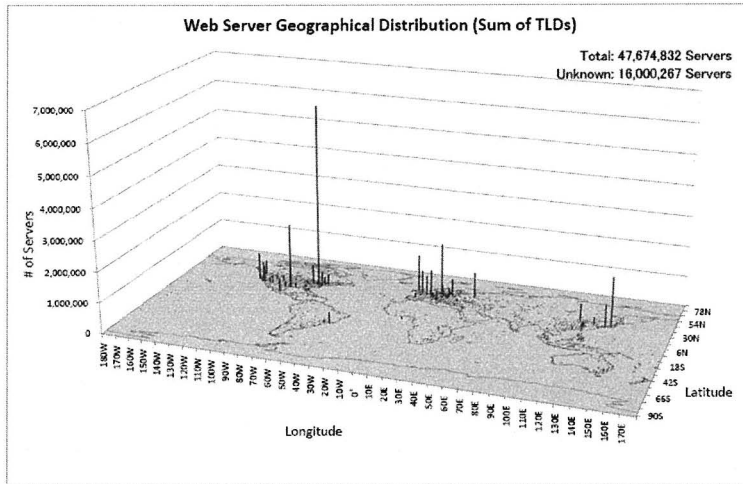


図 1 全世界の Web サーバの設置位置の分布

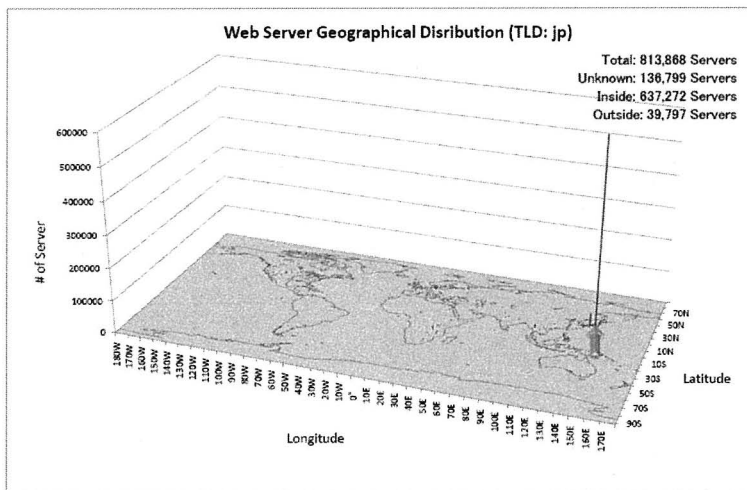


図 2 jp ドメインの Web サーバの設置位置分布

図 3 における 1 つのプロットは、1 つの ccTLD の値に対応する。さらに、グラフの左上にプロットされればされるほど、国外 Web サーバ数の割合が極端に強い ccTLD であり、右下にプロットされればされるほど国内に Web サーバ数の割合極端に強い ccTLD であることを意味する。ここで図 3 を見ると、3 つの傾向を見ることができる。

1. 国外 Web サーバ数の割合が極端に多い ccTLD は多数存在する。
2. 自国内 Web サーバ数の割合

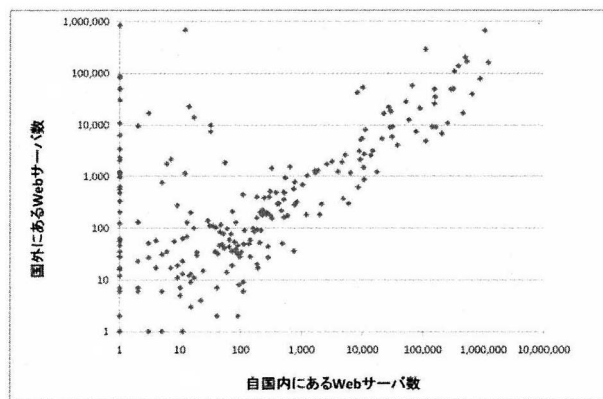


図 3 国内サーバ数・国外サーバ数の関係

が約 95%以上と極端に多い ccTLD は存在しない。

3. 自国内 Web サーバ数がある一定以上の値となると、自国内 Web サーバ数と国外 Web サーバ数に相関関係がある。

そこで、自国内 Web サーバ数が 60 以上<sup>h</sup>である ccTLD を対象を限定し、自国内 Web サーバ数と国外 Web サーバ数の関係の累乗近似式の算出を行った。その結果、自国内 Web サーバ数と国外 Web サーバ数の関係は、式(2)のような関係で近似できる。

$$y = 1.076x^{0.863} \quad (2)$$

ここで、 $y$  は国外 Web サーバ数を指し、 $x$  は自国内 Web サーバ数を指す。乗数が 0.863 であることから、自国内 Web サーバ数が多い ccTLD ほど、国外 Web サーバ数の割合が小さくなる傾向がある。

さらに、自国内 Web サーバ数に対する国外 Web サーバ数の適正範囲を、上記近似式を用いて定義を試みる。つまり、適正範囲を逸脱した ccTLD は、自国内 Web サーバに対して、国外 Web サーバ数が多すぎる、または少なすぎるといった理由で、異常 ccTLD として検出する。上記近似式を基準値とし、自国内 Web サーバ数に対する国外 Web サーバ数の適正範囲を仮に基準値の 1/10 倍以上、10 倍以下と設定すると、図 4 に示すように、自国内 Web サーバ数が 60 以上の ccTLD は、ほとんどが適正範囲内に位置する。そこで、自国内 Web サーバ数  $x$  に対する国外サーバ数  $y$  の適正範囲を式(3)のように定義する。

$$\text{適正範囲} : 0.1076x^{0.863} \leq y \leq 10.76x^{0.863} \quad (3)$$

この適正範囲外に該当する ccTLD が、自国内 Web サーバ数と国外 Web サーバ数の関係がアンバランスな ccTLD であると解釈することができる。

図 4 に適正範囲外の ccTLD に ccTLD 名を付与した。自国内サーバ数が多い ccTLD にも関わらず、国外 Web サーバ数が多いとして適正範囲外となった ccTLD は、be (ベルギー)、hr (クロアチア)、lu (ルクセンブルク) の 3 ドメインであった。また、国内 Web サーバ数が多いとして適正範囲外となった ccTLD は、bh (バーレーン)、bj (ベナン)、bt (ブータン) の 3 国であった。

#### 4.2 Web サーバの IP アドレス数分布

Web サーバとして利用されている IP アドレス(以下、IP アドレス数とする。)の地理的分布について解析を行った。本解析の目的は、同一ホストでの Virtual Host 数についての傾向を見い出すことにある。

約 4700 万 Web サーバの IP アドレスの経度・緯度情報を算出し、IP アドレスの地理的設置位置の特定を行った。表 4 に、世界の地域別の IP アドレス数、およびその割合を示す。Web サーバ数の割合と比べ

<sup>h</sup> 自国内に設置されている Web サーバ数の閾値を、10 から 200 の間で 10 ずつ変化させ、その都度、累乗近似式を計算した。閾値を 60 としたとき、累乗近似式の R<sup>2</sup> 乗値が最大となったため、閾値として 60 を選択した。

表 4 Web サーバ用 IP アドレスの地域分布

設置場所	IP アドレス数	割合
アジア	360,467	16.14%
アフリカ	8,867	0.40%
オセアニア	38,202	1.71%
中東	15,389	0.69%
中南米	46,198	2.07%
北米	1,088,797	48.76%
ヨーロッパ	674,513	30.21%
その他	361	0.02%

て、IP アドレス数の割合が増加した地域は、「アジア」「アフリカ」「オセアニア」「中東」であり、減少した地域は「中南米」「北米」「ヨーロッパ」である。

この結果から、Virtual Host を多く立てているホストは、「中南米」「北米」「ヨーロッパ」に多く存在することがうかがえる。

ここで、Web サーバ設置国別に Web サーバ数と IP アドレス数の関係をプロットすると、図 5 に示すように強い相関があることがわかる。そこで、4.1.2 節と同様に、IP アドレス数に対する Web サーバ数の適正範囲の定義を行い、適正範囲から外れた国を IP アドレス数に対して、Web サーバ数が多すぎる、少なすぎるといった理由で検出を行う。Web サーバ数  $y$  と IP アドレス数  $x$  の関係の近似式は式(4)に示す通りである。

$$y = 1.11x^{1.23} \quad (4)$$

上記近似式を基準値とし、IP アドレス数に対する Web サーバ数の適正範囲を基準値の 1/2 倍以上、2 倍以下と仮に設定すると、図 5 に示すように、ほとんどの国が適正範囲内にプロットされる。そこで、IP アドレス数  $x$  に対する Web サーバ数  $y$  の適正範囲を式(5)のように定義する。

$$\text{適正範囲} : 2.22x^{1.23} \leq y \leq 0.555x^{1.23} \quad (5)$$

IP アドレスと Web サーバ数の近似式、適正範囲および、適正範囲を逸脱した国の例を図 5 に示す。

図 5 を見ると、IP アドレス数に対して Web サーバ数が多いとして判定された国は、ロシア、スロベニア、ベラルーシ、マケドニア、グルジアの東欧諸国、ブラジル、メキシコの中南米諸国で多数を占められている。これらの国々に設置されているホストは、多数の Virtual Host を立てている傾向があると結論づけられる。

さらに詳しい分布を実行するために、図 6 では、経度・緯度 2 度四方のセル毎に、Web サーバ数基準値と比較して何倍の Web サーバ数が存在するかをグラフ化したものである。すべてのセルに対して、次に示す様な計算を実行した。

1. 当該セル内に設置されている IP アドレス数から Web サーバ数の基準値を計算
2. 実存していた Web サーバ数を基準値で除算

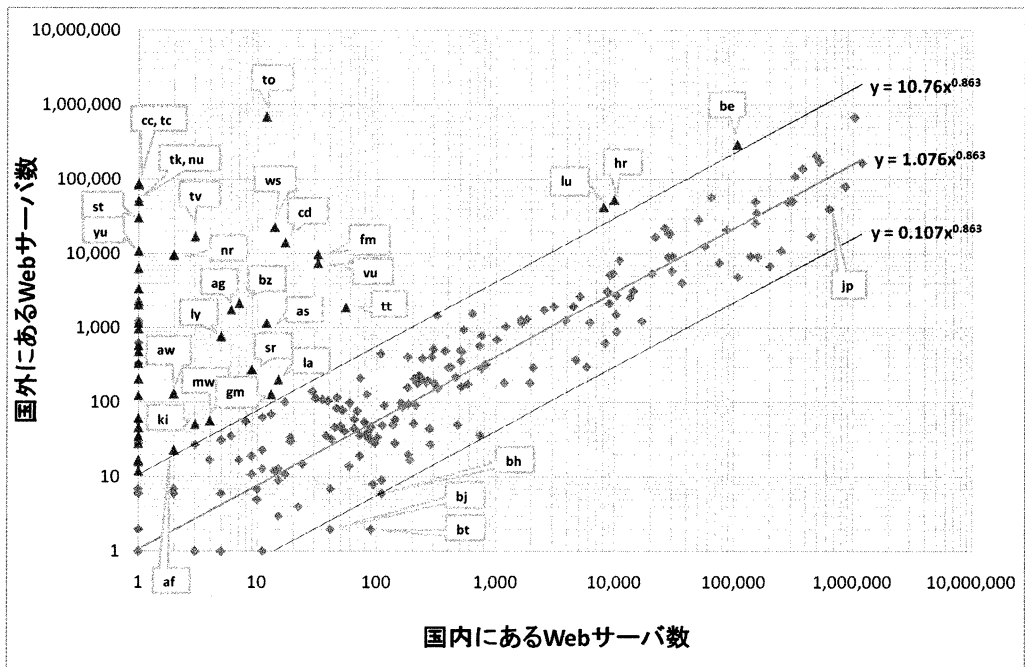


図 4 自国内 Web サーバ数と国外 Web サーバ数の適正範囲とアンバランスな ccTLD

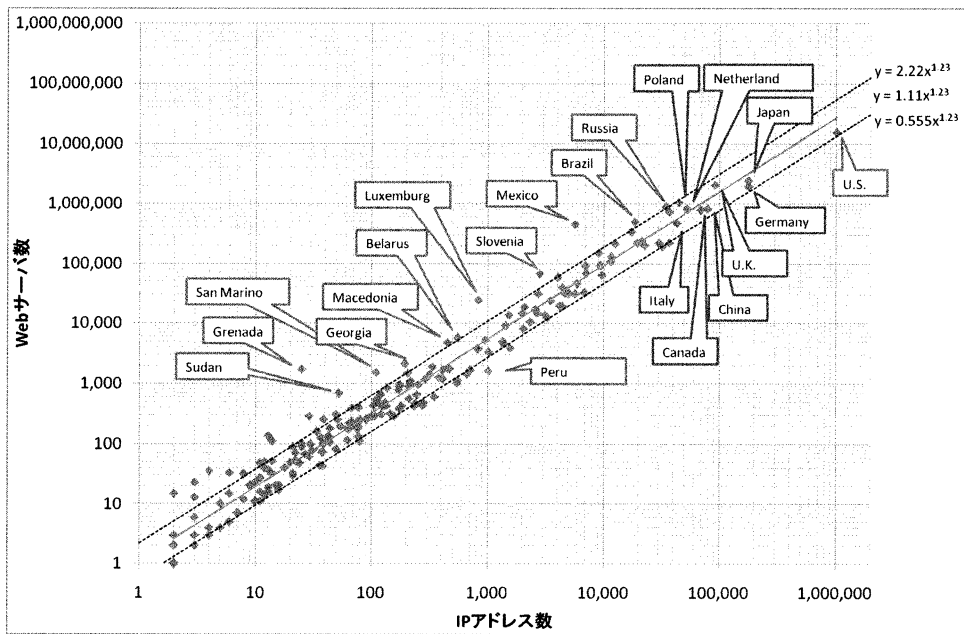


図 5 Web サーバ数と IP アドレス数の関係

図 6 に示す通り、ブラジル国内の西経 50 度、南緯 20 度付近のセルが基準値の約 130 倍の Web サーバ数が設置されており、この付近に設置されている Virtual Host 数が異常に多い結果となった。ブラジル

他には、アメリカ内陸部、ポーランド付近、メキシコシティにおいて、基準値と比較して大幅に多い Web サーバ数が検出された。

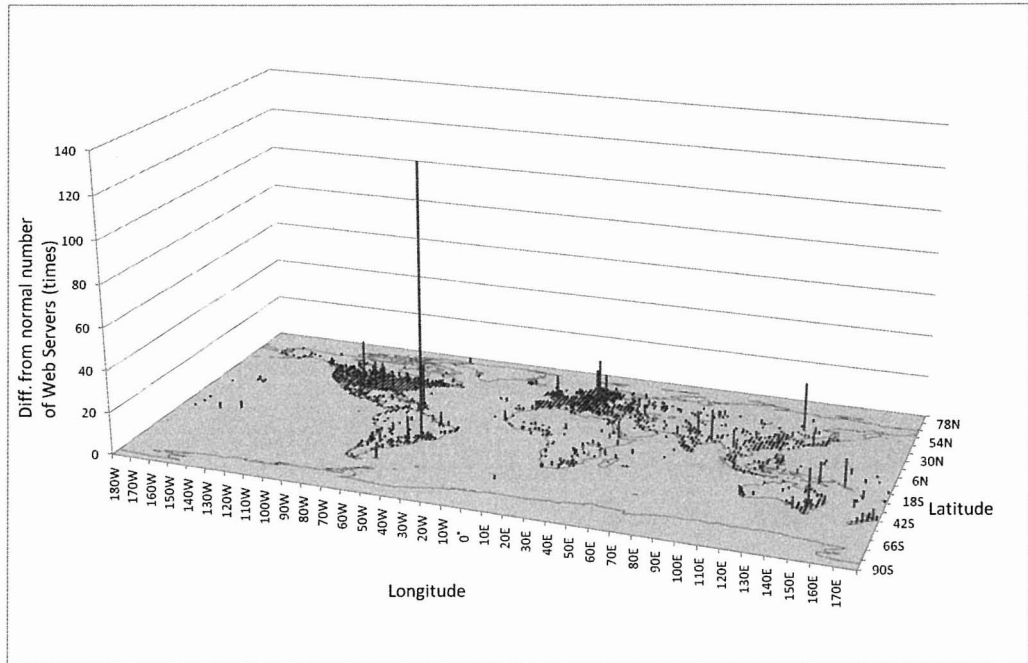


図 6 (実存した Web サーバ数 / Web サーバ数基準値) の値の分布  
 値が大きいセルほど Virtual Host 数が多い地域である。なお、値が 2 倍以上のセルを赤色のグラフで表現

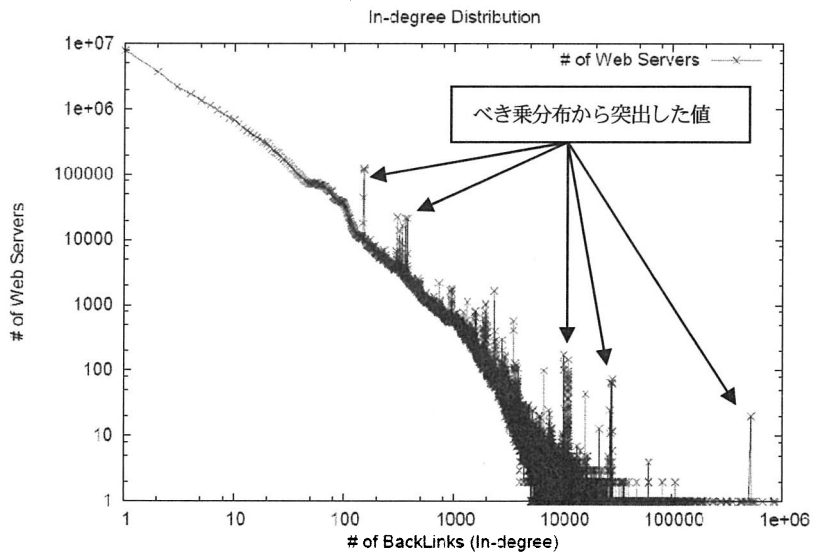


図 7 Web サーバの In-degree 値の分布

#### 4.3 バックリンク数分布

Web サーバのバックリンク数 (= In-degree 値) の分布状況を図 7 に示す。図 7 に示すように、In-degree 値と Web サーバ数の関係に Power-Low の関係が成立している。ここで、既存研究と同様に、In-degree 値  $x$  と

Web サーバ数  $y$  の関係式を示すと式(6)のようになる。

$$y = (1.5e + 7)x^{-1.45} \quad (6)$$

ここで、 $\gamma$  の値は 1.45 となり、既存研究の  $\gamma$  の値よりも小さな値となった。これは、我々の解析が Web ページ単位ではなく、Web サーバ単位であることに起因する

ると考えられる。すなわち、バックリンクを Web サーバ単位で集約を行っているため、In-degree 値が高い Web サーバが多く存在するためである。

また、図 7 の分布において、べき乗分布の関係から一部突出した値が存在している。これらの突出値は、同一バックリンク数(=In-degree 数)の Web サーバ数が、通常よりも突出して多いことを意味するので、スパム Web サーバである可能性が高いと考えられる。これらのべき乗分布からの突出した値に対応する Web サーバに関する具体的な調査は、今後の課題としたい。

## 5. おわりに

本稿では、e-Society プロジェクトで収集した約 107 億ページのページデータに対して、Web サーバの地理的設置位置の特定、およびバックリンク数 (=In-degree 数) の分布解析を行った。

その結果、北米・ヨーロッパ・アジア地域に設置されている Web サーバ数は、全体の約 95.5% であった。すなわち、中東、アフリカ、中南米地域を合計しても約 4.5% の割合にしかならず、現在の Web サーバの設置地域が偏っていることを示した。

さらに ccTLD ごとに、当該 ccTLD 管理国内に設置されている Web サーバ数と国外に設置されている Web サーバ数の関係を計算した。その結果、国内設置 Web サーバ数と国外設置 Web サーバ数に相関関係があることを発見した。さらに、国外設置 Web サーバ数が多いとして相関関係から外れた ccTLD は多数検出されたが、国外設置 Web サーバ数が少ないとして相関関係から外れた ccTLD はほとんど検出されなかった。これは、どの ccTLD においても例外なく、一定以上の割合で、ドメイン管理国外に Web サーバが設置されていることを意味している。

また 1 ホストあたりに立てられている Virtual Host 数が多いホストは、中南米、東欧諸国に多く見られ、特にブラジル・ポーランドに設置されているホストが顕著であった。また、米国の内陸部、メキシコシティに設置されているホストも Virtual Host 数が多いと検出された。

さらに、Web サーバのバックリンク数(In-degree 値)の分布について計測を行ったところ、既存研究と同様に、Power-Low の関係が成立していることを確認した。しかし、べき乗分布から外れた値が多数検出された。この外れ値は Spam サイトである確率が高いと考えられるため、今後は、べき乗分布から外れた値に対応する Web サーバを解析対象としたい。

## 謝辞

本研究の一部は、文部科学省リーディングプロジェクト「e-Society」及び情報爆発プロジェクトとして実施した。

## 参考文献

- 1) Hirate, Y., Kato, S. and Yamana, H.: “Web Structure in 2005”. In Proc. of WAW2006, LNCS 4936, pp. 36–46 (2008).
- 2) Kamvar, D. S., Haveliwala, T. H., Manning, C. and Golub, G.: “Exploiting the block structure of the web for computing PageRank”, Technical Report, Stanford University (2003).
- 3) Flake, G., Lawrence, S. and Giles, C.: “Efficient Identification of Web Communities”, In Proc. of 6th ACM SIGKDD Conf., pp.150-160 (2000).
- 4) 池内淳, 安形輝: “World Wide Web の測定: Web ページ推定手法の比較”, 三田図書館・情報学会 2001 年度研究大会(2001), [http://www.daito.ac.jp/~ikeuchi/webmetrics/webmetrics\\_1.html](http://www.daito.ac.jp/~ikeuchi/webmetrics/webmetrics_1.html)
- 5) Bar-Yossef, Z. And Gurevich, M.: “Random Sampling from a Search Engines’ s Index”, in Proc. of WWW2006, pp. 367–376 (2006).
- 6) e-Society プロジェクト, <http://www.yama.info.waseda.ac.jp/~yamana/es/>
- 7) 加藤真, 山名早人: “Fact of the Web:30 億ページのウェブの解析”, DEWS2006, 3B-i6 (2006).
- 8) 平手勇宇, 山名早人: “全世界の Web ページの TLD・言語分布解析”, 第 70 回 IPSJ 全国大会, 5L-1 (2008).
- 9) 童芳, 平手勇宇, 山名早人: “全世界の Web サイトの言語分布と日本語を含む Web サイトのリンク・地理的位置の解析”, DEWS2008, A2-3 (2008).
- 10) Nakahira, K., Hoshino, T. And Mikami, Y.: “Geographic Location of Web Servers under African Domains”, In Proc. of WWW2006, pp. 989 – 990 (2006).
- 11) 近藤浩之, 手塚太郎, 田中克己: “リンク元ページのアドレス情報に基づく Web ページの地域的支持度の分析”, DEWS2007, B9-4 (2007).
- 12) Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A. and Upfal, E.: “The Web as a Graph”, In Proc. of PODS2000, pp. 15–17 (2000).
- 13) Barabasi, A. and Albert, R.: “Emergence of Scaling in Random Networks,” Science, Vol. 286, pp. 509–512, (1999).
- 14) Broder, A., Kumar, R., maghoul, F., Raghavan, P. Rajagopalan, S., Stata, R. Tomkins, A. and Wiener, J.: “Graph Structure in the Web”, In Proc. of WWW2000, pp.309–320 (2000).
- 15) Boldi, P., Codenotti, B., Santini, M. and Vigna, S.: “Structural Properties of the African Web”, In Proc. of WWW2002, Poster Session (2002).
- 16) Liu, G., Yu, Y., Han, J. and Xue G.: “Gina Web Graph Measurements and Evolution”, In Proc. of APWeb2005, LNCS 3399, pp. 668–679 (2005).
- 17) IP2Location, IP-Country-Region-City-Latitude-Longitude-ISP-DomainDatabase[DB8], <http://www.ip2location.com/ip-country-region-city-latitude-longitude-isp-domain.aspx>