

全世界のWebサイトのTLD・言語分布・地理的設置位置の特定

Identifying Distributions of TLD, Written Language and Geographical Location of the Whole Web

童 芳[♥] 平手 勇宇[♦] 山名 早人[▲]

Hou TOU Yu HIRATE Hayato YAMANA

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。これまで、Webを対象としたリンク解析、テキスト解析等が幅広く行われているが、解析精度の向上のためには、大規模なWebページを対象とした統計情報が必要である。本論文では、2004年1月～2006年7月の間に収集した約107億Webページに対し、トップレベルドメイン(TLD)分布・言語分布の解析を行う共に、Webサーバの地理的な設置場所の特定を行った。その結果、「com」に属するWebサーバが全サーバの37.8%、英語で記述されたページが全Webページの42.6%、米国内に設置されているWebサーバが全Webサーバの48.4%を占めることがわかり、Webが「com」、英語、米国に偏っていることがわかった。

According to our investigation result in Oct. 2005, the number of web pages all over the world is estimated 53.7 billion. Although various web analysis methods, such as link analysis and text analysis, are being widely applied to web pages, a huge statistical analysis is indispensable to improve accuracy of such web analysis. In this paper, we show statistical web information, such as top level domain (TLD) distribution, written language distribution and distribution of geographical web server location, based on 10.7 billion Web pages. Our results show that 37.8% of web servers belong to "com" domain, 42.6% of pages are written in English and 48.4% of web servers are placed in United States. These results mean current web is highly biased to "com", "English" and "United States".

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[1]。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[2][3][4][5]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。

従来の研究・調査では、調査項目が限定的であるものが多い。International Software Consortium (ISC) [6]は、半年に1回、インターネットに接続しているホストのドメイン名の分布状況について統計調査を行っている。しかし、ISCでは、WebサーバやWebページの分布に関する調査は行われていない。Webサーバ数の推移については、Netcraft社¹が自社のホームページ上で毎月、Web Server Surveyとして発表している。Netcraft社は1995年から毎月調査結果を発表しており、インターネット上で唯一のWebサーバ数の統計情報となっている。しかし、Webサーバのドメイン別の分布やWebページ数に関する統計は存在しない。

Webページのドメイン分布については、池内ら[7]が検索エンジンのドメイン指定によるクエリを用いた統計的な調査を行っており、Altavista²を用いた2001年の調査では、com (48.0%)、net (7.8%)、edu (6.7%)、jp (4.8%)となっている。一方、ISCの統計調査における2001年1月時点のインターネットに接続するサーバ数の分布は、com (35.1%)、net (25.2%)、edu (7.2%)、jp (3.7%)となっており、池内らの調査と順序は等しいものの分布が大きく異なることがわかる。さらに、Netcraft社による2001年1月時点のWebサーバ数分布では、com (57.4%)となっているが、一部のTLDについてのみの情報公開となっており、全体像は不明である。

以上から、インターネットに接続するホスト数とWebサーバ数は必ずしも比例しないことがわかる。さらに、検索エンジンを用いて分布を求めた場合、その分布に誤差が生じることがわかる。これは、検索エンジンが持つインデックスに偏りが依存するためではないかと考えられる。

Webページの言語の分布についても、検索エンジンを用いた調査が2001年2月に池内らによって行われているが、ドメイン分布と同様、検索エンジンのインデックスに左右されるという問題を持つ。

Webサーバの設置位置に関しては、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、58.2%のWebサーバが米国内に設置されていたという報告[9]があるが、大規模に調査した例は著者の知る限り存在しない。

一方、検索エンジンを用いずに独自にWebページを収集し解析するという試みが行われている。2006年の著者らのメンバーである加藤らは、30億のWebページを対象として調査を行っている[9]。しかし、全Webページの6%程度を解析したに過ぎない。

以上示したように、これまでの研究・調査は、調査項目が限定的であったり、調査規模が小さい。これに対し、本論文では、e-Societyプロジェクト[10]によって2004年1月から2006年7月までに全世界から収集した107億Webページに対し、WebページのTLD分布、WebサーバのTLD分布、Webページの言語分布、およびWebサーバの地理的設置場所の分布の解析を行った。

本論文では、次のような構成をとる。まず、第2節では既存の統計情報と地理位置の関連研究について述べる。次に、第3節でデータセットと具体的な解析方法について述べる。第4節では各統計結果を示し、第5節でまとめる。

[♥] 学生会員 早稲田大学大学院理工学研究科
ygmstf@yama.info.waseda.ac.jp

[♦] 正会員 早稲田大学メディアネットワークセンター
hirate@yama.info.waseda.ac.jp

[▲] 正会員 早稲田大学理工学術院
yamana@yama.info.waseda.ac.jp

¹ <http://survey.netcraft.com/Reports/>

² <http://www.altavista.com/>

2. 関連研究

2.1 Webサーバ・WebページのTLD分布に関する研究

2001年2月に、池内らはISCの調査結果に対して検索エンジンを用いて、Webページ数と言語分布について統計調査を行った[7]。その結果によると、Webページ数の分布は、ISCが統計したホストのドメイン分布結果とほぼ一致したが、一部のドメインに分布の偏りが存在したと報告している。ホスト数に比べてWebページ数の比率が多いドメインとしてcom, org, de, gov, ruなどが挙げられた。一方、ホスト数が多いにも関わらず、Webページ数が少ないドメインとしてnet, milなどがあった。jpドメインの場合は、ホスト数とWebページ数のいずれの結果でも、全体の3~4%を占めると報告されている。

検索エンジン用いたTLD分布状況の研究として、2006年に行われたBar-Yossefらの研究[11]がある。Bar-Yossefらは、検索エンジンのインデックスをランダムサンプリングする手法を提案しており、[11]において検索エンジンのインデックスサンプルからTLDの分布状況について言及している。その結果、検索エンジンとしてGoogle, MSN, Yahoo!のどの3社を選択したとしても、comのドメインの割合はおおよそ50%前後である結果となっている。しかし、第1節で言及したように、検索エンジンのインデックスに依存する問題点をもつ。

これに対し著者らのメンバーである加藤らは、2004年1月から2005年7月末までに収集したWebページ中の約30億のWebページに対してWeb構造を中心に解析を行った[9]。しかし、ここでの解析対象は主としてcomドメインとjpドメインであり、Webサーバ数分布は、com (76.9%)、jp (22.3%)であり、Webページ数分布は、com (86.2%)、jp (11.5%) となるなど、全世界のWebページの統計としては利用できない。

2.2 Webサーバの言語分布に関する研究

池内らの研究[7]では、検索エンジンを用いてWebページの言語分布についても調査を行っている。[7]では、英語ページの割合は、全Webページの約60%であると報告している。ただし、検索エンジンを用いて分布を計算している特性上、検索エンジンのインデックスに影響してしまう問題点がある。

また、加藤らも約30億ページの言語分布について解析を行っており、約66.4%が英語ページ、約13.2%が日本語ページであったと報告している[9]。しかし、対象としたWebページがcom, jpに偏っているため、英語、日本語のページの割合が大きくなっていると考えられる。

2.3 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[12]。[12]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[8]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。[8]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。

IP2Location社は、国別のIPアドレス所有数分布に関する統計情報を公開している[13]。[13]では、2007年の時点で、約37.7%のIPアドレスが米国に、約12.8%が英国に、約7.6%が日本に所有されていることを報告している。しかし、IPアドレス空間の割合であり、Webサーバの設置国分布に適用することはできない。

2.4 本論文の位置づけ

以上示したように、これまでの研究・調査は、調査項目が限定的であったり、調査規模が小さい。これに対し、本論文では、2004年1月から2006年7月までに全世界から収集した107億Webページに対し、WebページのTLD分布、WebサーバのTLD分布、Webページの言語分布、およびWebサーバの地理的設置場所の分布の解析を行う。

3. 解析方法

3.1 解析対象となるWebページ

解析対象のWebページは、e-Societyプロジェクト[10]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社が発表している統計情報によれば、全世界のWebサーバ数³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⁴を対象として解析を行った。

3.2 解析方法

3.2.1 解析対象とするWebページ情報

Webの傾向調査のために、全てのWebページに対し、次の3つの情報を抽出し、解析の対象とした。

TLD情報

IANAX[14]が2007年に公開した標準となる272個のTLD名に

³ 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くな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。

⁴ 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。

よって、収集されたページが属するTLDを判定した。なお、ドメイン名が割り当てられていないIPアドレスについては、「その他」として分類を行った。

言語情報

Webページの記述言語について、ベイシスの言語判定システム[15]を利用し言語判定を行った。判定対象言語は、英語、日本語、中国語、フランス語、韓国語、スペイン語、ドイツ語、イタリア語、ロシア語、ポルトガル語、アラビア語の12言語である。また、画像、動画等は、バイナリとして判別される。12言語に判別できなかった言語は、その他「oth」として分類される。

経度・緯度情報

任意のWebページの地理的位置を特定するために、まずWebページからWebページが配信されているWebサーバのIPアドレスをDNS参照により求めた⁵。次に、IP2Location社のIP-経度・緯度-国名・市名変換テーブル[16]を用いて、IPアドレスから経度・緯度情報への変換を行った。

3.2.2 Web解析項目

Webの統計情報を明らかにするために、次に示す4つの項目について解析を行った。

Webページ・WebサーバのTLD分布

解析対象の107億ページのページ毎、Webサーバ毎のTLD分布について調査を行った。

全Webページの言語分布・TLD毎の言語分布

解析対象の107億ページの言語分布の調査を行った。また、任意のTLDに含まれるWebページが、どの言語で書かれているのかを把握するために、TLD毎の言語分布の調査を行った。

全Webサーバの地理的設置位置分布・TLD毎の地理的位置分布

全世界を、経度2度、緯度2度四方セル⁶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。

4. 解析結果

4.1 Webページ・WebサーバのTLD分布

解析対象となる107億ページに対し、TLD毎のWebページ数、TLD毎のWebサーバ数（ホスト数）を調査した。図1に結果を示す。また、TLD毎のWebページ数分布を図2に示す。図1、図2では、Webサーバ数が多い上位20位までのTLDと、21位以下のTLDを“other”としてまとめあげた際の分布を示した。

表1は、TLD毎のWebページ数の分布とWebサーバ当たりの平均Webページ数を示す。表1に示されるように、ドメインによってWebサーバ当たりの平均Webページ数が大きく異なることがわかる。従来、1台あたりの平均Webページ数は200ページ前後と考えられていたが[2][3]、日本やドイツでは、これを大幅に上回るページ数が確認された。主な原因は、バックエンドにDBを持ち各種情報を提供するサーバが両国で急増しているためだと考えられる。

4.2 言語分布

107億ページの言語の分布状況を図3示す。英語が全体の

⁵ 複数のIPアドレスが返ってきた場合、1番目のIPアドレスを当該WebサーバのIPアドレスとした。

⁶ 経度2度緯度2度四方の領域は、おおよそ関東甲信越地方の大きさと同等である。

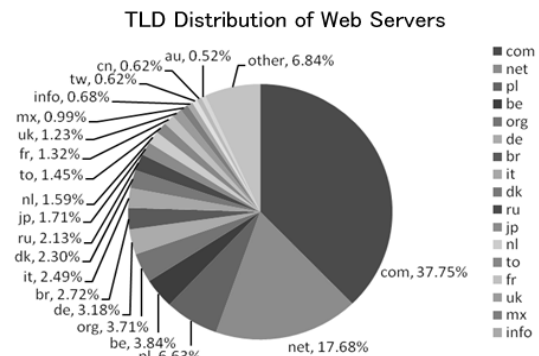


図1 107億ページに対応するWebサーバのTLD分布

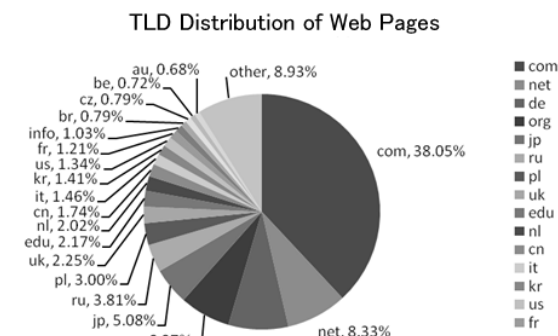


図2 107億ページのTLD分布

表1 TLD毎のWebページ分布

TLD	国名	取得ページ数	割合	平均ページ/サーバ
.com	-	4,070,092,124	38.05%	339
.net	-	890,604,259	8.33%	169
.de	ドイツ	878,838,449	8.22%	459
.org	-	745,984,032	6.97%	(※)
.jp	日本	543,654,556	5.08%	609
.ru	ロシア	407,169,769	3.81%	(※)
.pl	ポーランド	321,209,334	3.00%	185
.uk	イギリス	240,244,507	2.25%	(.com, .net, .jp, .de 以外の平均)
.edu	-	232,132,978	2.17%	
.nl	オランダ	215,722,380	2.02%	
.cn	中国	185,907,711	1.74%	
.it	イタリア	156,657,707	1.46%	
.kr	韓国	151,025,640	1.41%	
.us	アメリカ	143,135,686	1.34%	
.fr	フランス	129,326,495	1.21%	
other	-	1,385,290,926	12.95%	

42.6%を占め第一言語となっていることが分かる。一方で日本語はWeb上で第二言語となっており13.0%を占める。第三言語は独語の9.8%となっている。

図4はjpドメインに属するWebページの言語分布を表したものである。jpドメインから収集したページは、543,654,556ページであり、その中の90.3%が日本語で記述されたページである。英語で記述されたページは5.9%である。日本語、英語以外の言語で記述されたページはどれも1%未満となっている。

また、gTLDである「com」、 「net」及び、言語判定対象の

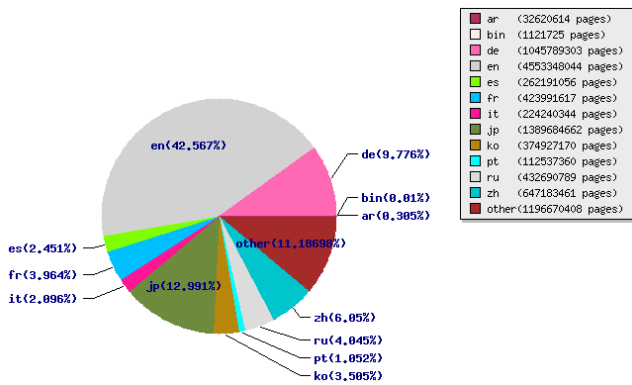


図 3 全世界 107 億ページの言語分布

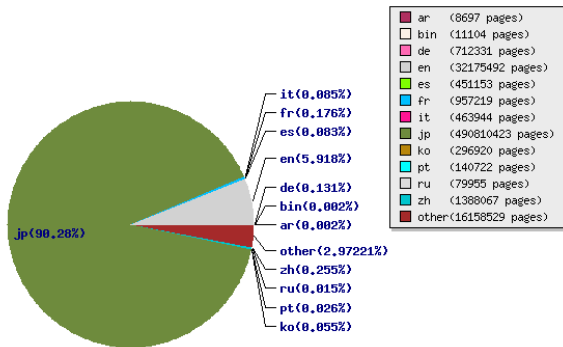


図 4 jp ドメインに属する約 5.4 億ページの言語分布

言語を母国語とする国のccTLDの言語分布を表 2に示す。gTLDのcomドメインとnetドメインにおいては、英語の割合がそれぞれ47.8%, 55.7%と高い割合となっている。これに対し、言語判定対象の言語を母国語とする国のccTLDは、jpドメインと同様に、当該言語の割合が高い結果となった。これは、当該国民に対して情報が発信されていることを意味する。

ただし、esドメイン（スペイン）におけるスペイン語の割合、及び、pt（ポルトガル）におけるポルトガル語の割合が、それぞれ62.6%, 52.3%となっており、表 2に掲載した他のccTLDの母国語の割合と比べて相対的に低い値となっている。

表 2 com, net 及び言語判定対象の言語を母国語とする国の ccTLD の言語分布

国名	ar	bin	de	en	es	fr	it	jp	ko	pt	ru	zh	その他	収集ページ数	公用語	
.com	0.55%	0.00%	2.97%	47.77%	3.82%	5.10%	1.29%	18.29%	4.46%	0.51%	0.43%	8.55%	6.27%	4,070,092,124		
.net	0.59%	0.00%	4.10%	55.67%	1.79%	3.36%	2.16%	13.03%	3.64%	0.38%	0.99%	8.18%	6.10%	890,604,259		
.de	ドイツ	0.03%	0.01%	85.50%	10.04%	0.25%	0.45%	0.37%	0.06%	0.02%	0.09%	0.17%	0.08%	2.93%	878,838,449	de
.es	スペイン	0.01%	0.05%	0.74%	62.63%	1.23%	0.76%	0.20%	0.03%	1.23%	0.05%	0.14%	8.67%	28,442,593	jp	
.fr	フランス	0.01%	0.03%	0.62%	16.71%	0.55%	76.18%	0.46%	0.08%	0.01%	0.17%	0.05%	0.08%	5.03%	129,326,495	fr
.it	イタリア	0.06%	0.04%	1.59%	15.82%	0.61%	0.68%	78.48%	0.06%	0.02%	0.20%	0.04%	0.15%	2.24%	156,657,707	it
.kr	韓国	0.00%	0.00%	0.06%	4.21%	0.02%	0.19%	0.04%	0.08%	94.14%	0.01%	0.01%	0.08%	1.15%	151,025,640	kr
.pt	ポルトガル	0.01%	0.03%	0.91%	37.07%	1.03%	1.03%	0.47%	0.20%	0.09%	52.31%	0.26%	0.09%	6.54%	7,321,422	pt
.ru	ロシア	0.01%	0.02%	0.25%	9.62%	0.07%	0.15%	0.08%	0.02%	0.01%	0.03%	87.70%	0.03%	2.00%	407,169,769	ru
.cn	中国	0.11%	0.00%	0.11%	5.15%	0.11%	0.15%	0.12%	0.27%	0.07%	0.01%	0.12%	91.16%	2.62%	185,907,711	zh

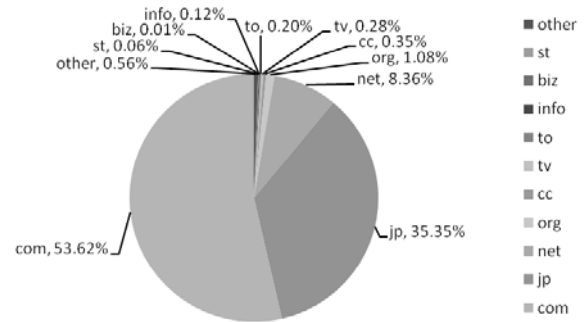


図 5 日本語で書かれたページの TLD 分布

その分、es ドメイン、pt ドメインにおける英語の割合が高くなっていることは、注目に値する。

4.3 日本語 Web ページの TLD 分布

次に、日本語で記述されたWebページのTLD分布について解析した。対象データは107億のWebページであり、この中から日本語で記述されていると判定された約13%のページのTLD分布を図 5に示す。図 5に示すように、jpドメインよりもgTLDであるcomドメインの方が割合が高くなっている。jpドメインは35.4%でcomドメインに次いで2位となった。結論として、日本語Webページの半分以上は地域の明示されていないgTLDに属しており、約1/3がjpドメインに存在することが分かる。この結果は、日本語のページの収集を目的として、jpドメインのみを収集対象としても、日本語のページの約3分の1しか収集できないことを意味している。

4.4 Web サーバの地理的設置位置分布

4.4.1 全Webサーバの地理的設置位置分布

107億ページを収集した約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 6に示す。図 6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としては、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。

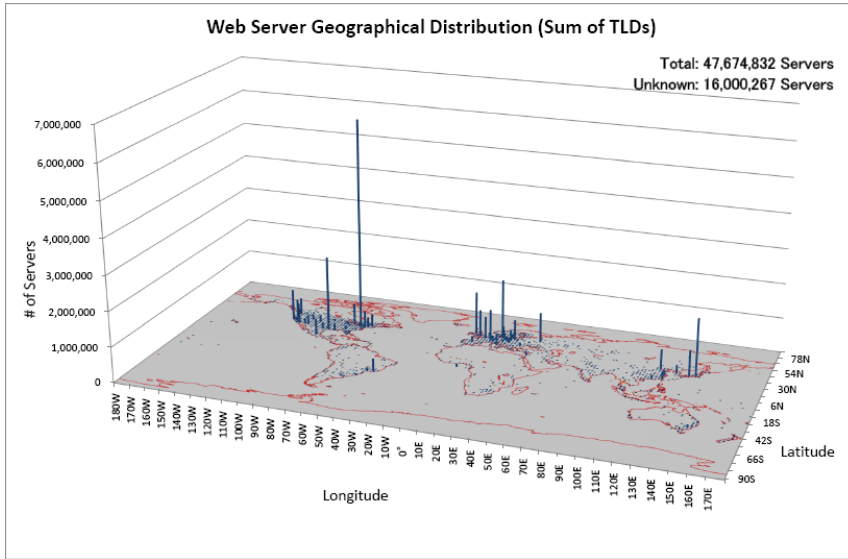


図 6 全世界のWebサーバ設置位置分布

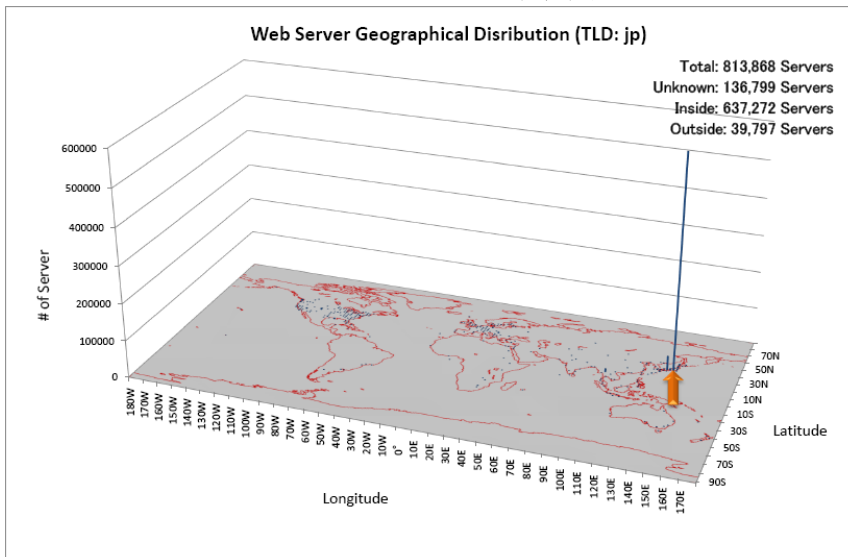


図 7 jpドメインに属するWebサーバの設置位置分布

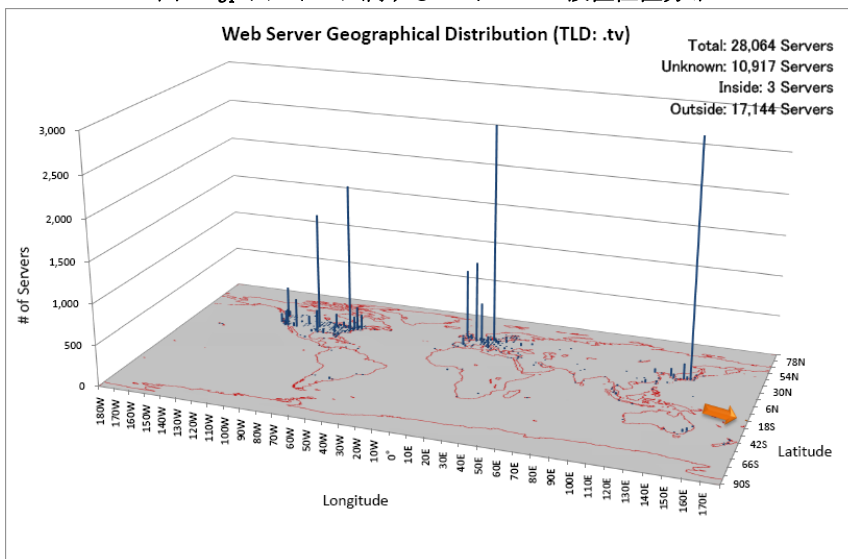


図 8 tvドメインに属するWebサーバの設置位置分布

表 3 全世界のWebサーバ設置国分布

設置場所	Webサーバ数	割合
United States	15,325,293	48.38%
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%

表 4 jpドメインに属するWebサーバの設置国分布

設置場所	Webサーバ数	割合
Japan	637,272	94.12%
United States	19,962	2.95%
Thailand	9,746	1.44%
Hungary	2,599	0.38%
Australia	1,288	0.19%
Germany	992	0.15%
Taiwan	745	0.11%
China	534	0.08%
Indonesia	464	0.07%
Korea	377	0.06%
United Kingdom	362	0.05%
Nepal	332	0.05%
Malaysia	321	0.05%

表 5 tvドメインに属するWebサーバの設置国分布

設置場所	Webサーバ数	割合
United States	7,091	41.35%
Japan	3,168	18.48%
Germany	2,791	16.28%
France	1,021	5.95%
United Kingdom	955	5.57%
Netherland	464	2.71%
Canada	289	1.69%
Italy	188	1.10%
Korea	146	0.85%
Spain	124	0.72%
Australia	81	0.47%
China	70	0.41%
Switzerland	56	0.33%

図 6に示す通り、Webサーバの大部分が、米国及び欧州に設置されていることがわかる。具体的には、米国に設置されているWebサーバ数は15,325,293台となっており、設置位置が特定できた全Webサーバ31,674,565台の内、48.4%を占める。日本国内に設置されていたWebサーバは、2,437,179台であり、全体の7.7%となった。Webサーバ設置数が多かった国の上位14ヶ国を表 3に示す。

4.4.2 jpドメインに属するWebサーバの地理的設置位置分布

図 7は、jpドメインに属するWebサーバの設置位置の分布を示した図である。図 7右上の「Inside」の数字は、日本国内に設置されていたWebサーバ数を示し、「Outside」の数字は日本国外に設置されていたWebサーバ数を示す⁷。

jpドメインに属するWebサーバは、米国、欧州に設置されているものも存在するが、設置位置が特定できたWebサーバのうち、94.1%が日本国内に設置されていることが確認できた。なお、2番目に設置が多かった国は米国であり、jpドメインに属するWebサーバのうち、3.0%が設置されている結果となった。表 4に、jpドメインに属するWebサーバが設置されている割合が多い13ヶ国を示す。

4.4.3 tvドメインに属するWebサーバの地理的設置位置分布

図 8は、tvドメインに属するWebサーバの設置位置の分布を示した図である。tvはツバルが管理するccTLDであり、ツバルは太平洋の中に存在する島国である。図 7に示すように、tvに属するWebサーバの内、ツバル国内に設置されているWebサーバは、わずか3台であった。その他のWebサーバは、表 5に示すように、米国、日本、独国等のインターネット普及率が高い国に設置されている結果となった。

5. おわりに

本稿では、2004年1月から2006年7月までに収集された世界中の107億Webページに対して、TLD分布、言語分布、Webサーバの地理的設置位置分布の調査を行った。その結果、現在のWebは、「comドメイン」、「英語」、「米国」の3つに偏っていることが判明した。

また、jpドメインに属するWebサーバの地理的設置位置の特定を行った結果、94.1%が日本国内に設置されていることが判明した反面、tvドメインのようにドメインを所有している国の国内にはほとんどWebサーバが設置されていないドメインも存在していることも判明した。一方で、日本語で記述されたWebページの内、約3分の2が、jpドメイン外から配信されていることがわかった。このことは、日本語ページに限定したWebページの収集や解析が、jpドメインのみを対象としたのでは不十分であることを示す。

【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。

【文献】

- [1] Hirate, Y., Kato, S. and Yamana, H.: "Web Structure in 2005". In Proc. of WAW2006, LNCS 4936, pp. 36-46 (2008).

⁷ すべての TLD に対して図 8, 図 9 のような分布図を生成しているその他のドメインについては、<http://www.yama.info.waseda.ac.jp/e-society/>にて公開している。

- [2] Lawrence, S. and Giles, C. L.: "Searching the World Wide Web", Science, Vol.280, No.5360, pp.98-100 (1998).
- [3] Lawrence, S. and Giles, C. L.: "Accessibility of Information on the Web", Nature, Vol.400, pp.107-109 (1999).
- [4] 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).
- [5] Flake, G., Lawrence, S. and Giles, C.: "Efficient Identification of Web Communities", In Proc. of 6th ACM SIGKDD Conf., pp.150-160 (2000).
- [6] Internet Software Consortium, <http://www.isc.org/>
- [7] 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
- [8] 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).
- [9] 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).
- [10] e-Society プロジェクト, <http://www.yama.info.waseda.ac.jp/~yamana/es/>
- [11] Bar-Yossef, Z. And Gurevich, M.: "Random Sampling from a Search Engines's Index", in Proc. of WWW2006, pp. 367-376 (2006).
- [12] Nakahira, K., Hoshino, T. And Mikami, Y.: "Geographic Location of Web Servers under African Domains", In Proc. of WWW2006, pp. 989-990 (2006).
- [13] IP2Location(TM) Internet IP Address 2008 Report, <http://www.ip2location.com/ip2location-internet-ip-address-2008-report.aspx>.
- [14] IANA: Root-Zone Whois Index by TLD Code, <http://www.iana.org/root-whois/index.html>
- [15] Basis Technology Rosette 言語判別システム, <http://www.basistech.co.jp/language-identification/>
- [16] IP2Location
IP-Country-Region-City-Latitude-Longitude-ISP-Do
mainDatabase[DB8], <http://www.ip2location.com/ip-country-region-city-latitude-longitude-isp-domain.aspx>

童 芳 Hou TOU

2008 早大・理工学研究科修士課程修了。現在、TIS 株式会社に勤務。

平手 勇宇 Yu HIRATE

2005 早大・理工学研究科修士課程修了。2008 早大・理工学研究科博士課程修了。博士(工学)。2006 年より同大・メディアネットワークセンター助手。ACM, IEEE, IPSJ, DBSJ 各会員。

山名 早人 Hayato YAMANA

1993早大・理工学研究科博士課程了。博士(工学)。1993-2000 電総研。2000早大・理工助教授。2005同大理工学術院教授、現在に至る。IEEE, ACM, IEICE, IPSJ, DBSJ各会員。