全世界のWebページのTLD・言語分布解析

平手 勇宇[†] 山名 早人[‡]

早稲田大学大学院理工学研究科[†] 早稲田大学理工学術院[‡]

1.はじめに

我々は，インターネット情報知識集約を可能にするプラットフォーム構築を目指し，2004年1月下旬から2006年7月までの約30ヶ月の期間に，約144億ページのWebページの収集を行った[1]．収集した全世界のWebページ解析をするにあたり，ページデータセットの全体傾向を把握することは，Webページ解析結果を向上させるために重要である．この目的のため，我々は収集したWebページの統計的な調査を行っている．本稿では，データとして利用可能な107億ページ[2]のWebページを対象とし，トップレベルドメイン(以下TLDとする)，記述されている言語の2つの切り口で調査を行った結果を報告する．

2.107億ページのTLD・言語分布状況

2004年1月9日から2006年7月31日の期間に収集された約144億ページの内，10,696,996,553ページに対して解析を行った．なお，Webページの収集の際に，Basis TechnologyのRosette言語判定システム[3]を利用し，当該Webページがどのような言語で記述されているのかの言語判定を行っている[1]．本節では，約107億ページのTLD分布，言語分布について示す．

収集した約107億ページのTLD分布を表1に示す．全世界には合計272個存在する[2]が，表1では，取得ページ数が多かったTLD上位15個，および16位以下のTLDの合計をOtherとしてまとめ上げた．表1より，".com"に属するWebページは，全体の38.05%を占めており，Webが".com"に偏っていることがわかる．また，com,net,org等のgTLDのドメインに属するウェブページの割合は全体の56.95%と，全体の半分以上を占めている．

また，国別コードトップレベルドメイン(ccTLD)の中で，一番大きな割合となったのはドイツ(8.22%)であった．TLD上位15位中のccTLDが対応する国は，ロシア・ポーランド・中国を除き，全てインターネット普及率が2007年6月現在50%以上[3]であった．

図1に収集した107億ページの言語分布を示す．なお図1の判例のbinは，画像等のバイナリであることを意味する[4]．図1に示す通り，英語で書かれたページが，全体の42.57%を占めている．次に多かった言語は日本語の13.00%であるが，これは我々が日本語ページを起点としてWebページを収集したため，我々のWebページクローリングが，日本語に偏っていることが原因であると考えられる．

表1：全収集ページのTLD分布

TLD	国名(ccTLDのみ)	取得ページ数	割合
.com	-	4,070,092,124	38.05%
.net	-	890,604,259	8.33%
.de	ドイツ	878,838,449	8.22%
.org	-	745,984,032	6.97%
.jp	日本	543,654,556	5.08%
.ru	ロシア	407,169,769	3.81%
.pl	ポーランド	321,209,334	3.00%
.uk	イギリス	240,244,507	2.25%
.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%

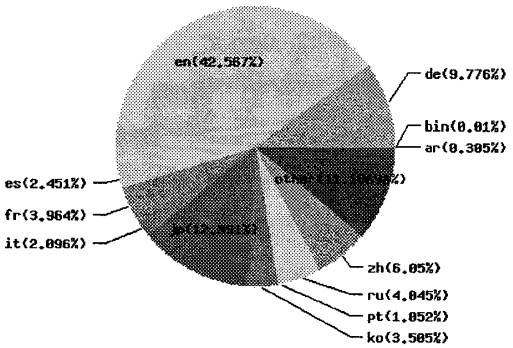

図1：全収集ページの言語分布

3. TLD毎の言語分布解析

次に，収集したページをTLDごとに分類し，TLDごとに言語割合を計算し，当該国のインターネット普及率[3]と比較する[5]ことで，TLD毎の言語分布解析を行った．紙面の都合上，全TLDの結果を示すことができないため，本稿では一部のTLDのみ記載を行う[6]．

表2では，TLDごとの収集したWebページの言語分布，およびインターネット普及率を示したものである．.comドメイン，.netドメインの2つのTLDは，一般TLD(=gTLD)でかつWebページ数が多いため，Web全体の言語分布と似た言語分布となる結果であった．

---

[1] 2007年9月以降は，日本語Webページを含むホストを対象に，更新Webページの収集を目的として，取得済みのホストに含まれるWebページの再収集を行っている．

[2] 37億ページはDISK故障等によりデータが失われた．

[3] アラビア語(ar)，ドイツ語(de)，英語(en)，スペイン語(es)，フランス語(fr)，イタリア語(it)，日本語(jp)，韓国語(kr)，ポルトガル語(pt)，ロシア語(ru)，中国語(zh)，バイナリ(bin)，その他に判別することが可能である．

[4] 我々は，バイナリを収集しないポリシーでクローリングを行ったため，実際のバイナリの割合はもっと大きな値となる．

[5] ccTLDのみを対象とした．

[6] 全ドメインの言語分布結果は，
http://www.yama.info.waseda.ac.jp/e-society/を参照のこと．

表 2: ccTLD ごとの言語分布率，収集ページ数，公用語およびインターネット普及率

ccTLD	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	普及率
com	-	0.55%	0.00%	2.97%	47.77%	3.82%	5.10%	1.29%	18.29%	4.46%	0.61%	0.43%	8.55%	6.27%	4,070,092,124		
net	-	0.59%	0.00%	4.10%	56.67%	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	64.6%
jp	日本	0.00%	0.00%	0.13%	5.92%	0.08%	0.18%	0.09%	90.28%	0.05%	0.03%	0.01%	0.26%	2.97%	543,654,556	jp	68.0%
ch	スイス	0.04%	0.01%	51.12%	29.41%	0.56%	11.41%	2.58%	0.11%	0.02%	0.16%	0.14%	0.24%	4.19%	55,544,767	de, fr, it	69.2%
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	57.0%
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	67.1%
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	54.7%
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	19.5%
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	12.3%
to	トンガ	0.00%	0.00%	2.90%	17.37%	0.14%	1.54%	0.66%	59.20%	0.64%	0.04%	0.53%	14.77%	2.22%	4,623,520	トンガ語 en	3.0%
dj	ジブチ	0.03%	0.00%	75.91%	11.24%	0.25%	6.31%	0.14%	0.14%	1.85%	0.02%	3.57%	0.00%	0.54%	211,451	fr	1.4%
cc	ココス諸島	0.26%	0.00%	5.76%	57.84%	0.61%	0.76%	2.88%	20.00%	1.89%	0.10%	0.24%	7.57%	2.12%	24,599,794	en	n/a

3.1 インターネット普及率が高い国

インターネット普及率が 50%以上と高い国は全部で 37 カ国存在する．これらの国が該当する ccTLD の全 Web ページのうち，おおよそ 7 割以上が母国語で書かれていた [4]．

例外は，スイス，ガージン島の 2 カ国である．スイスは公用語でない英語ページが全体の 29.41%を占めており，公用語のフランス語，イタリア語よりも大きな割合を占めていた．ガージン島は，フランスに隣接する小さな島国であり，公用語がフランス語・英語であるにもかかわらず，ドイツ語の Web ページの割合が，25.5%を占めていた．これは，ガージン島が国策としてドメインを売っており，ドイツのサイトが gg ドメインを利用していることが考えられる．

3.2 インターネット普及率が低い国

インターネット普及率が低い国々には，3.1 で示したガージン島と同じ現象が多く見られる．たとえば，トンガ(.to ドメイン)の公用語は，トンガ語・英語であるにも関わらず，.to ドメインに属する Web ページのうち，日本語のページが 59.2%の割合であった．また，ジブチ(.dj ドメイン)では，ドイツ語が 75.91%も占めており，これはジブチ在住の人を対象にしたページとは考えにくい．ココス諸島 (.cc ドメイン) に関しても，公用語でない日本語のページが目立つ[7]．

このような現象は，インターネット普及率が低く，かつ面積が非常に小さな国で頻発する．このほかにも，アセンション島(.ac)，ベリーズ(.bz)，クリスマス諸島(.cx)，ミクロネシア連邦(.fm)，グレネダ(.gd)，サウスジョージア・サウスサンドウィッチ諸島(.gs)，イギリス領インド洋地域(.io)，プエルトリコ(.pr)，，セントヘレナ(.sh)，サントメ・プリンシペ(.st)，ツバル(.tv)等が挙げられる．

[7] ココス諸島では，公用語ではないが中国語も利用されている [4]．

4.おわりに

本稿では，全世界の Web ページ約 107 億ページを対象として，Web 全体の傾向調査，および TLD，言語の二つの切り口から，Web の傾向の調査を行った．

Web 全体では，.com に属する Web ページが全体の 38.05%と偏っていた．また，国別に，Web ページの言語分布状況を調べた結果，インターネット普及率が高い国々では，おおよそ当該公用語で書かれた Web ページが大きな割合を占める結果であった．しかし，普及率が小さく，面積が小さな国々では，ドメインを売る政策の影響のため，公用語でない言語で書かれた Web ページが多数存在している結果となった．

参考文献

[1] Basis Technology, Rosette 言語判定システム, http://www.basistech.co.jp/language-identification/

[2] IANA, http://www.iana.org/

[3] Internet Usage World Stats - Internet and Population Statistics, http://www.internetworldstat.com/

[4] CIA - The World Factbook, https://www.cia.gov/library/publications/the-world-factbook/