D-047

ccTLDを単位としたWebコミュニティ構造の分析

Analysis of Web Communities grouping by ccTLD

石原　直幸†　　中平　勝子†　　三上　喜貴†
Naoyuki Ishihara　Katsuko T. Nakahira　Yoshiki Mikami

1．はじめに

Web 空間のグラフとしての構造については，従来様々な研究が行われてきた．Broder らのパイオニア的分析[1]は Web 空間全体を幾つかの成分に分別し，「蝶ネクタイ」と名付けられているマクロ構造があることを明らかにした．またミクロレベルでは，Kumar らが Web 空間には無数の完全二部グラフ構造があるとして，ハブとオーソリティの概念を抽出した[2].

他方，グラフ構造における各ノードの次数分布パターンについても，多くの分析や観察が行われてきた．そこでは各 Web ページに張られるリンクの数はべき乗則に従う分布になることが明らかにされている[1][3].

ウェブコミュニティの定義は様々であるが，本研究では国を単位とする Web サイトの集合を国単位の巨大なコミュニティと定義し，そのグラフ構造を分析しようとするものである．特に Web コミュニティの内部構造について，外向きリンクの次数（out-degree）の分布パターン（out-degree 分布図）が各国の Web コミュニティの成熟度を示す「診断チャート」として活用できるのではないかと考え，その有効性について，情報先進国や発展途上国を含むアジア 42 ヶ国を対象として分析を行った．また国単位の Web コミュニティ相互間のリンク構造についても分析した.

2．分析方法

2.1　データの取得方法

本研究で用いたデータは，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]を用いて収集した．クローリングを行った期間は 2006 年 7 月 5 日から 2006 年 7 月 19 日までの 15 日間であり，対象としたカントリーコード(ccTLD，以降同じ)はアジア地域の 42 ヶ国（日本，韓国，中国を除く）の ccTLD である．クローリングは 13,286 個のシード URL から開始した．またクローリングは以下の条件の下で行った.

1) ネットワークの回線容量やディスクスペースの容量のため，クローリングを行う深さを 8 階層まで，1 つのホストからの収集ページ数を 5 万ページまでに制限
2) robots.txt が置かれている Web サイトを回避
3) 収集するファイルは HTML 及び TXT ファイルに限定

このクローリングにより取得できたデータ量は 652,710,237,381 バイト，Web ページ数は 107,168,782 ページ，Web ページから張られているリンクの総数は 2,974,522,875 個であった．本研究ではこれらの Web ページの URL とその Web ページから張られているリンク先 URL のデータを利用した.

なお，国の単位の Web コミュニティをどのように把握するかについては議論のあるところだが，本稿においては ccTLD により国の判別を行った．またジェネリック・トップレベルドメイン(gTLD，以降同じ)は，その帰属する国の特定を行うことが困難であるため調査対象から除いた.

2.2　ccTLD別のページ数とリンク数

クローリングを行った国と ccTLD，収集できた Web ページ数，Web ページから張られているリンク総数を表 1 に示す．最も多くの Web ページを有するのはイスラエルで約 3000 万ページに及び，最も少ない国は東ティモールの約 1 万 3000 ページであった．両国間には約 2300 倍もの差がある.

2.3　out-degree 分布図

指標として用いた out-degree とは，グラフ理論における用語であり，有向グラフにおいて 1 つの頂点から出ていく枝の数を表す．Web 空間を巨大な有向グラフと見なせば，これは 1 つの Web ページ上から他の Web ページに対して張られているリンクの総数を表すものである．本研究で使用する out-degree 分布図とは，out-degree の値を横軸にとり，縦軸にはその out-degree を持つページ総数をとった両対数グラフのことである．この out-degree 分布図の形状は，過去の研究から，ほぼべき乗則に従うことが示されている[1][3]．本稿では，この手法をアジア地域 42 ヶ国の国別 Web コミュニティに対するクローリングデータの分析に適用した.

また，Web ページに対して張られているリンクは，同じ Web サイト内へのリンク，異なる Web サイトへのリンクと 2 種類に分けることができ，それぞれに対しての out-degree 分布図を作成した.

2.4　TLD間のリンク強度

国間の Web コミュニティの繋がり強度を調べるために，調査対象となる国の ccTLD を持つ Web ページから張られているリンクの数をリンク先のウェブページの TLD 別に集計し，TLD 間のリンク強度マトリックスを作成した.

3．結果と考察

3.1　Web利用の拡大と out-degree 分布の変化

アジア 42 ヶ国それぞれに対して out-degree 分布図を作成した．その内からページ数の順位において，均等な間隔で 6 ヶ国を抽出して示したのが図 1 である．図 1 の分布

† 長岡技術科学大学

表 1　ccTLD 別の取得データ量

Country	ccTLD	ページ数	リンク数
＊東ティモール	tp	13,213	262,090
ミャンマー	mm	16,759	771,127
パプアニューギニア	pg	27,011	3,377,888
イエメン	ye	34,128	1,268,256
モルディブ	mv	37,393	482,748
ブータン	bt	44,594	1,005,000
シリア	sy	51,555	1,121,179
カタール	qa	52,888	3,041,656
＊クウェート	kw	59,152	876,923
カンボジア	kh	64,265	623,351
トルクメニスタン	tm	80,509	3,491,113
パレスチナ	ps	88,203	3,052,061
ブルネイ	bn	94,788	537,467
スリランカ	lk	136,519	1,713,345
アフガニスタン	af	141,263	3,211,764
オマーン	om	145,207	1,956,644
＊ラオス	la	146,635	9,733,940
バングラデシュ	bd	207,150	6,381,511
タジキスタン	tj	233,623	3,518,616
バーレーン	bh	246,031	9,011,208
ヨルダン	jo	287,341	8,384,882
レバノン	lb	343,538	5,059,425
ネパール	np	395,901	12,224,659
モンゴル	mn	400,141	11,360,654
＊キプロス	cy	627,056	14,761,185
パキスタン	pk	734,989	13,034,732
キルギスタン	kg	740,921	21,653,169
アラブ首長国連邦	ae	934,634	28,654,246
サウジアラビア	sa	1,053,672	166,702,006
アゼルバイジャン	az	2,251,487	63,620,374
ウズベキスタン	uz	2,286,738	69,648,109
フィリピン	ph	2,732,525	73,147,485
＊イラン	ir	4,022,272	89,297,141
インド	in	4,262,379	100,983,106
ベトナム	vn	4,490,292	144,990,287
インドネシア	id	5,742,097	201,799,579
シンガポール	sg	5,771,198	120,881,989
カザフスタン	kz	6,441,381	156,701,462
マレーシア	my	6,865,807	177,223,036
トルコ	tr	11,363,643	304,036,842
＊タイ	th	12,556,823	332,057,717
イスラエル	il	30,943,061	802,862,903

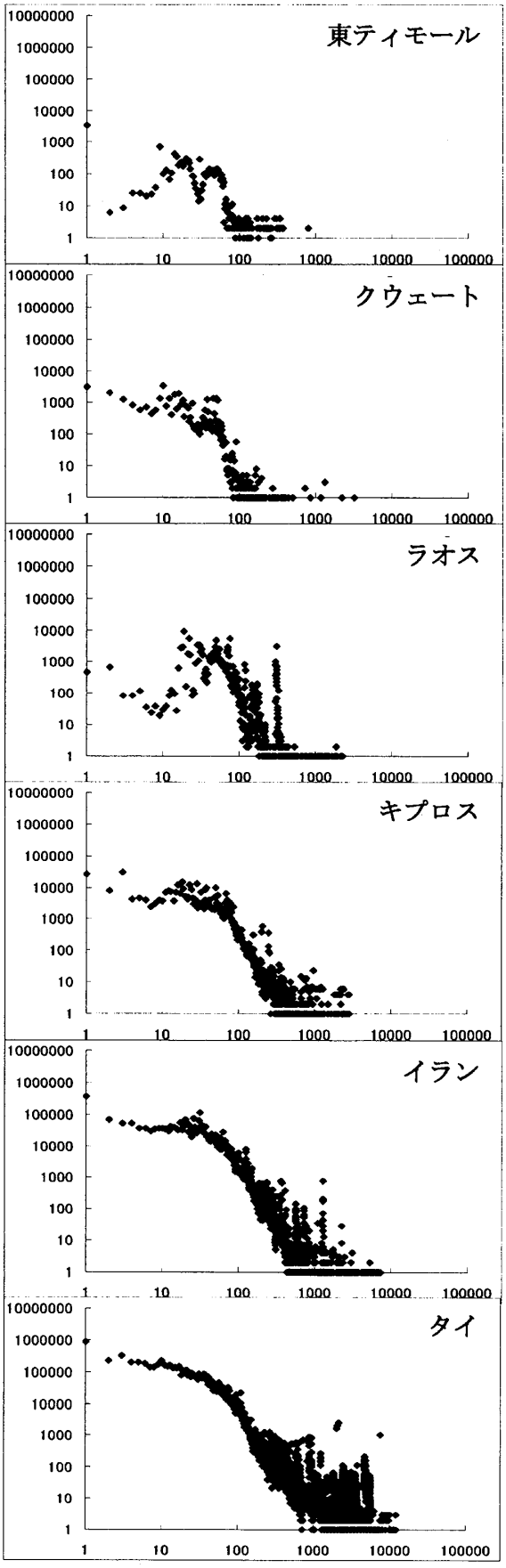

図 1　Web ページ数順に並べた out-degree 分布図
x 軸は out-degree，y 軸はページ数を示す

図では，クローリングにより得たデータをリンク先により分割したり，動的生成ページを除くなど特別な処理を行わず示したものである．作成したグラフ群から以下の 3 点の特徴を見いだした．

1. Web ページの総数が増加するに従い，グラフは整然としない形から，まとまりを持つ形へと変化していく
2. out-degree が 30 程度まではほぼ水平を示す
3. それを超えるとべき乗則に従う直線を描く

以上に加え，out-degree が 1000 を超えるような非常に大きな値では，垂直線状に伸びた形が現れる国もいくつか存在した．

特に注目したいのが上記 1 であり，Web ページ数が十分に増えれば共通の形に収束していくことが確認できた．これは，多数の当事者が制約を受けずに Web 空間上の活動を展開する時，一定の out-degree パターンに近づくことを意味していると解釈できる．Barabási らは，べき乗則が成立するのは，

1) 新しい頂点（ページ）の追加によってネットワークが継続的に成長していること
2) 新しく追加された頂点が多数の枝（リンク）を持つ頂点と繋がりやすい選択傾向を持つこと

の 2 つの条件が満たされるであると述べている[5]．今回の調査で out-degree 分布図が変則的な形状を示した国は，こうした Web 空間の自由な成長やリンク先選択が制限されている国ではないかと推測できる．実際に今回調査した国の中でも Web ページ総数が少ない東ティモール，ミャンマー，パプアニューギニアは他国と比較しても分布図の形が極めていびつであり，この 3 ヶ国に対して詳しく調査を行ってみると以下のように極めて少数のしかも突出した当事者の支配的な Web コミュニティであることが明かになった．

東ティモール：収集できた Web ページの約 97%が google ドメインによる Web ページである

ミャンマー：唯一の ISP は政府が担っており，インターネットカフェにて閲覧した Web ページが監視されるなど規制が行われている

パプアニューギニア：Web ページの 65%が政府機関に属するものである

こうした傾向が普遍的に成立するとすれば，out-degree 分布図により Web コミュニティの相対的な成熟度の判別が行えるのではないかと考えられる．

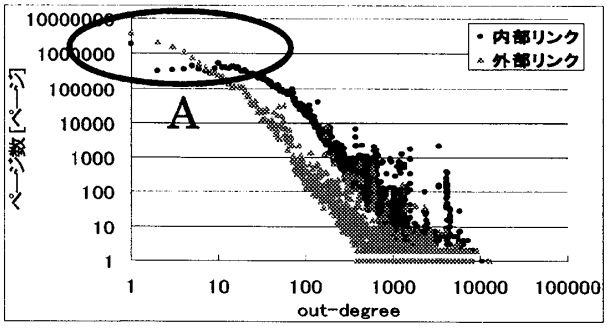

図 2　イスラエルの外部リンク・内部リンク分布図

3.2　内部リンクと外部リンク

アジア 42 ヶ国の全ての国に対してリンク先により内部リンク，外部リンクに分類した out-degree 分布図を作成したが，代表例としてイスラエルを示したのが図 2 である．図 2 において内部リンクが外部リンクより多く存在していることが分かるが，out-degree が 10 を下回る辺りでは外部リンクの数が逆転する（図 2 の A）．この部分は注目すべき形状を示している．図 1 のイスラエルの分布図では，out-degree が 30 を下回ると水平を示していたが，これは同じサイト内へのリンクである内部リンクが大きく影響していることが図 2 の A の部分より分かる．水平を示す理由としては，個々の Web サイトが巨大化してきたことにより，内部へのリンク数が 2 個や 3 個といった少数のリンク数では収まりきらず，10 個を超えるものが一般化してきていることが原因ではないかと考えられる．また図 2 の A の部分を除けば 2 つの分布はべき乗則に従い，ほぼ相似形となった．この 2 つの分布の out-degree の差は約 3 倍であり，このことは異なる Web サイトのページよりも同じ Web サイト内のページとの結びつきが 3 倍程度強いことを示しているのではないかと推測することができる．ここではイスラエルを例としてとりあげたが，Web ページが比較的多く存在する国々においても同様に 2 つの分布図が相似形を示す結果となった．

3.3　動的生成ページ

42 ヶ国に対して作成した out-degree 分布上では，垂直線状の形が，Web ページ総数が比較的に多い国々を中心にいくつか見られた．ラオスやネパールといった中規模程度の Web ページ総数を持つ国においては，ただ 1 つの垂直線が出現しているのに対して，イランやタイといった多数の Web ページを持つ国では 1 つに限らず多数の垂直線の形が現れ五月雨状となった．out-degree 分布図において，この垂直状の形が現れるということは，近い数のリンク数を持つ Web ページが極めて多く存在することを示している．これは，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，掲示板やスパム類のページなどの存在を示す痕跡であると考えられる．このことを裏付ける実例が，図 3 のネパールにおける PHP を用いて作成されたある掲示板を含む Web サイトのみを抜き出した out-degree 分布図である．同図(B)は，掲示板のような Web サイトが実際に垂直状の形を示すことを表している．他にマレーシアにおいて，大量の同一内容のスパム広告にリンクが張られている掲示板も同様な垂直線を示していることを確認した．

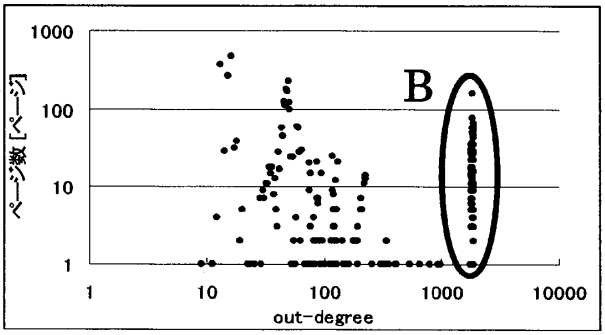

図 3　掲示板を含む Web サイトの out-degree 分布図

表2 TLD間のリンク強度

アフガニスタン		ラオス		モンゴル		トルコ		タイ		イスラエル	
af	68.5%	la	75.1%	mn	78.6%	tr	94.1%	th	91.3%	il	93.8%
com	22.0%	com	18.8%	info	12.8%	com	3.5%	com	4.3%	com	3.4%
de	4.0%	org	2.1%	com	4.3%	org	0.9%	net	1.0%	info	0.7%
org	3.4%	net	1.0%	be	2.3%	net	0.7%	info	0.9%	org	0.6%
sc	0.6%	cn	0.3%	net	0.4%	de	0.1%	org	0.9%	net	0.6%

3.4 TLD間のリンク強度

調査対象の国として，アジアの国々の中でも中規模のWebページ数を持つ3ヶ国（アフガニスタン，ラオス，モンゴル）と多数のWebページ数を持つ3ヶ国（トルコ，タイ，イスラエル）を対象として調査を行った．各国から張られているリンクをリンク先のTLD別に集計し，総数に対する比率を求めた．その上位5位までを示した結果が表2である．各国とも自国へのリンク率が最も高い割合を示す結果となった．しかしながら，その割合については Webページ総数の規模により差が見られた．中規模の3ヶ国においては 60~80%の割合であったが，大規模な3ヶ国においてはいずれも 90%以上の高い値を示した．これは，自国が持つ Webページ数が多ければ（コミュニティの規模が大きければ）より自己完結的な Webコミュニティの形成が可能となるからであると考えられる．これに対してそれほど Webページが多くない国では，自己完結的な Webコミュニティの形成が難しいことが原因であると考えられる．また調査を行った6ヶ国全ての国において，gTLDへのリンク率が自国以外のccTLDと比較して多い割合であった．これはgTLDはどの国の組織でも取得できるために自国に関連する Webサイトが含まれていたり，世界的に有名なニュースサイトなどが存在することが原因であると考えられるが今後追跡調査が必要である．

3.5 ccTLDによる国判別の妥当性

今回収集した Webページがどの国に属するものであるのかの判別はccTLDにより行った．しかし，"tv"など国外組織に販売されている ccTLDの存在[6]や，国外に置かれまたその管理も国外の専門家に任されている事例も報告されているため [7]，ccTLDによる国の判別を行うのでは正確な結果を得ることができない可能性も考えられる．そこで ccTLDを利用した判別方法で正確な結果を得ることが可能であるのか追加調査を行った．方法は"whois"コマンドを使用して，対象となる Webページのドメインを保有している組織ないし個人の登録住所を調べた．本稿では，成熟度の異なる3ヶ国（サウジアラビア，カザフスタン，イラン）について調査を行った．

結果を表3に示す．イランは判別が可能であった46個のドメイン全てがイラン国内の組織または人間により登録されているものであった．これはイランが日本の“jp”ドメインと同じように，ドメインを取得できる条件がイランに居住する者に限定されるなどてしていることが原因であると考えられる．また，サウジアラビア，カザフスタンもイランに準ずるように極めて高い割合を示した．なおカザフスタンの保有者が判明したドメイン数が，調査対象のドメイン数に対して極めて少ないのはカザフスタンの Whoisサーバに使用制限が存在したため全てを調査することが困難であったことが原因である．また，国によって結果は大きく異なることが予想されるので対象となる国全てに対しての調査が今後必要である．

表3 ドメインの登録組織の調査

	サウジアラビア	カザフスタン	イラン
調査対象のドメイン数	86	2530	67
保有者が判明したドメイン数	86	193	46
保有者が居住者のドメイン数	85	187	46
居住者の保有割合[%]	98.8	96.9	100.0

4．まとめ

本研究は，out-degree分布図によって，国を単位としたWebコミュニティの成熟度をそのグラフの内部構造の視覚化表現として示すことができるのではないかと考え，発展段階の異なる42ヶ国のデータを用いて out-degree分布図を作成し，相対的な成熟度を判断することが可能であることを示した．また自国内へのリンクと自国外へのリンクの比率にも成熟度が反映されていること，掲示板などの存在は同図の上で特徴的なパターンとして現れることを示した．さらに ccTLDによる国判断には妥当性があることを確認した．

5．参考文献

[1] A. Broder, R. Kumar, F. Maghoul et. al. : Graph structure in the web, The 9th International World Wide Web Conference, May 2000

[2] R. Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins: Trawling the web for emerging cyber communities, The 8th International World Wide Web Conference, May 1999

[3] D. Gleich, L. Zhukov: Scalable Computing for Power Law Graphs: Experience with Parallel PageRank, The International Conference for High Performance Computing, Networking, Storage and Analysis, November 2005

[4] Ubicrawler http://law.dsi.unimi.it/

[5] Albert-László Barabási, Réka Albert: Emergence of Scaling in Random Networks, Science 286, pp.509-512, October 1999

[6] 和田祥太: 島嶼国 ccTLDの有効活用と管理改善のための提言, 長岡技術科学大学大学院　工学研究科　修士課程修士論文

[7] 星野哲哉, 中平勝子, 三上喜貴: アジア・アフリカドメインにおけるサーバ地理的所在，電子情報通信学会信越支部大会講演論文集,pp. 153, 2006