Fact of the Web
－30 億ページのウェブの解析－

加藤真† 山名早人††,†††

†早稲田大学大学院理工学研究科 〒169-85555 東京都新宿区大久保 3-4-1
††早稲田大学院理工学術院 〒169-85555 東京都新宿区大久保 3-4-1

†††国立情報学研究所 〒101-8430 東京都千代田区一ツ橋 2-1-2
E-mail: †kato@yama.info.waseda.ac.jp, ††yamana@waseda.jp

あらまし 現在ウェブサーバから発信されている情報量は，2006 年 2 月の時点で静的なページだけでも 150 億ページ以上あると推測される．これは，全世界の Web ページ数を推定した過去の 3 つの研究から 1 ウェブサーバ当たりの平均ウェブページ数を 200 ページと仮定し，ウェブサーバ総数 7618 万台との積をとったものである．しかし，我々が 2006 年 2 月までに収集をした 120 億ページを元に推定すると 2006 年 2 月時点で 350 億ページが存在するという結果を得た．これは，近年動的に生成されるウェブページが急増していることに起因するものと考えられる．また，本稿では，これまでに収集した 120 億ページの内，30 億ページについてウェブ構造を中心に様々な解析を行った．その結果，最近のウェブ構造は、いわゆる「蝶ネクタイ」構造の真ん中（CORE）の部分が巨大化していることが判明した．特に，中国語や日本語でこの傾向が強いことがわかった．

キーワード Web とインターネット，知識発見，データマイニング

Fact of the Web
－Analysis of 3 Billion Web Pages－

Shin KATO† and Hayato YAMANA††,†††

†Graduate School of Science and Enginnering, Waseda University 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555 Japan
††Science and Engineering, Waseda University 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555 Japan

†††National Institute of Infomatices 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430 Japan
E-mail: †kato@yama.info.waseda.ac.jp, ††yamana@waseda.jp

Abstract The number of static web pages is estimated over 15 billion in Feb 2006. This is multiplying 200 pages by 76.18 million web servers, where 200 pages means the average number of web pages and are assumed from past three researches. However, based on the analysis of 12 billion web pages that we have crawled by Feb. 2006, we estimate the total number of web pages as 35 billion. This is because dynamic web pages are rapidly increased in recent years. And we also analyzed web structure using 3 billion web pages. As a result, we figure out that the size of "CORE", the center component of bow-tie structure, is increasing in recent years, especially in Chinese and Japanese web.

Key words Web and Internet, Knowledge Discovery, Data Mining

1. はじめに

現在ウェブサーバから発信されている情報量は静的なページだけでも，2006 年 2 月時点で，約 150 億ページと推測される．

ウェブの規模については，NEC 北米研究所の主任研究員であった Lawrence らによる推計 が有名である．1998 年に Science 誌に掲載された論文[1]では，同一の検索語を用いて複数の検索エンジンで検索し，複数の検索エンジンの検索結果の重なりの度合いからウェブページ数を推計している．推計にあたっては，各検索エンジンがインデックスしているページ数と検索結果の重なりの度合いを用いる．本手法によれば，1997 年末時点のウェブページ数は 3.2 億ページである．さらに，1999 年に Nature 誌に掲載された Lawrence らの論文[2]では，360 万個の IP アドレスに対して 80 番ポートでウェブサーバが立ち上がっているかを調査すると共に，2500 台のウェブサーバに対して実際にウェブページ収集を行っている．これにより，IP アドレス空間上の推定ウェブサーバ数とウェブサーバ当たりの平均ウェブページ数との積を求め，1999 年 2 月時点で約 8 億の Web ページが存在すると推計している．

文献[1]での 1 ウェブサーバ当たりの平均ウェブページ数は 190 ページ，文献[2]での 1 サーバ当たりの平均ウェブページ数は 186 ページであり，推計時期によらずほぼ一定となっている．また，総務省情報通信研究

所が 2004 年 2 月に実施した WWW コンテンツ統計調査[3]においても 1 ホスト当たりの平均ウェブページ数は 202 ページとなっていることから，現時点においても 1 サーバ当たりの平均 Web ページ数は，200 ページ前後であると推測できる．

以上の結果から，1 サーバ当たりのウェブページ数を 200 と仮定し,2006 年 2 月時点のウェブサーバ数 76,184,000[4]との積をとることにより約 150 億ページと推定できる．

しかし，我々が 2006 年 2 月までに収集した 120 億ページを元に全世界のウェブページ数を推定すると 2006 年 2 月時点で静的・動的ページを含めて 350 億と予想される．これは，近年 CGI 等によって生成される動的な Web ページが急増していることに起因すると予想される．

さらに，ウェブの構造に着目するとこれまでに様々な研究が成されている．1999 年に行われた Broder らによる研究[5]では，ウェブのページとリンクをグラフの頂点と辺とみなすと，全体の約 3 割のページが一つの強連結成分を成すと共に，約 9 割のページが一つの連結成分を成すことが報告されている．この連結成分の構造が，模式的に「蝶ネクタイ」の形を成していたため，「蝶ネクタイ構造」と呼ばれている．

一方，2002 年に行われた Boldi らによるアフリカのウェブの解析[6]においては，蝶ネクタイ構造を確認することができず，最大の強連結成分を中心とし，そこから他の複数の強連結成分へ連結するような構造となっていることが報告されている．また，2003 年に行われた Lie らによる中国のウェブの解析[7]においては，蝶ネクタイ構造の構成成分が 1999 年の解析結果と異なり，約 8 割のページが一つの強連結成分を成していることが報告されている．

これらの調査に対し，本研究では，最新のウェブページを用いた各種統計的な解析を行い最新の調査結果を示す，具体的には，e-Society プロジェクト[8]によって,全世界のウェブを対象に 2006 年 2 月末までに収集完了した 120 億ページの内、30 億ページを対象に解析を行った．

以下 2 節では全世界のウェブページ総数について述べ，以降はウェブページの解析について述べる．具体的には，3 節でウェブ構造関連研究，4 節では解析対象となるデータセットについて述べる．5 節では解析プラットフォームについて，6 節で統計情報について述べる．7 節で強連結成分の解析について述べ，8 節でまとめる．

2. 全世界のウェブページ総数

我々は，2006 年 2 月の時点で既知のホスト数が 60,968,174 ホストで,約 42%のホストの収集が完了し，12,003,683,320 ページを集めた．

このことから，1 サーバあたりのウェブページ数は，12,003,683,320 ÷ (60,968,174 × 0.42) から約 465 ページとなる．

[4]によると，2006 年 2 月時点のウェブサーバ数は 76,184,000 推定されており，この積をとることで，全世界のウェブページ総数は，静的・動的ページを含めて，約 350 億ページと予想される．これは，近年増加する．近年ブログ，ポータルサイトまたは，EC サイトの増加によって，動的に生成される Web ページが急増していることに起因すると予想される．

3. ウェブ構造関連研究

3.1. Graph Structure in the web[5]

Broder らは，1999 年に収集した約 2 億ページ，約 15 億リンクについて解析を行っている．この解析よると，ウェブ全体をグラフとして捉えると，図 1 のような蝶ネクタイ構造を成しており，ウェブページの約 9 割がひとつの連結成分を成している．

また,この連結成分は **CORE**, **IN**, **OUT**, **TENDRILS**, 4 つに分類できる．**CORE** はひとつの巨大な強連結なページ群，**IN** は **CORE** へは辿れるが，**CORE** からは辿れないページ群，逆に，**OUT** は **CORE** からは辿れるが，**CORE** へは辿れないページ群，**TENDRILS** は **CORE** から辿ることも，**CORE** へ辿ることもできないページ群である．

1999 年に収集したデータでは表 1 に示すとおり，**CORE** が全体の 3 割を占め，**IN**, **OUT**, **TENDRILS** はそれぞれ 2 割であった．

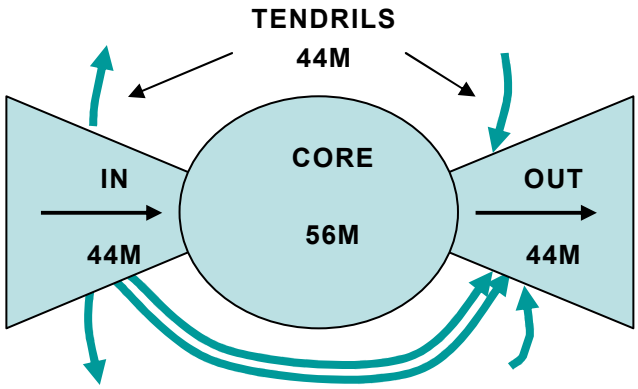

図 1蝶ネクタイ構造

3.2. Structural Properties of the African Web[6]

Boldi らは，2002 年 2 月に収集した 200 万ページ，2500 ホストのアフリカのウェブページに関して解析を行っている．

この解析では,Border らの解析結果と異なり,**CORE** は存在したが，**CORE** から辿れるが，**CORE** へは辿れない，**CORE** より小さい強連結成分がいくつか存在した．

3.3. China Web Graph Measurements and Evolution [7]

Lie らは，2003 年の 5 月に収集した，約 1.4 億ページ，約 43 億リンクについて解析を行っている．

この解析の結果，中国のウェブ構造は蝶ネクタイ構造を成していたが，CORE が特に巨大化し，約 8 割のページが強連結成分を成しており，中国特有の構造であると述べられている．

表 1 1999 年と 2003 年のウェブ構造の成分

Web Graph	CORE	IN	OUT	TENDRILS	DISCONNECT
1999 年の Web[5]	0.56 億 約 28%	0.43 億 約 21%	0.43 億 約 21%	0.44 億 約 22%	0.17 億 約 8%
2003 年の China Web [7]	1.1 億 約 81%	0.17 億 約 12%	0.09 億 約 6%	合計 0.01 億 約 1%	

4. 対象とするウェブページ

我々が収集した総ウェブページは，2004 年 1 月から 2006 年 2 月末までで 120 億ページである．本稿では、解析にあたり 2005 年 7 月末までに収集した 7,050,571,172 ページ中の 3,935,592,289 ページ，3,652,232 ホストについて解析を行った．なお、言語判定を用いた解析では，言語判定を適用した 3,193,373,141 ページを対象とした．

収集にあたっては，2005 年 7 月までは，収集ロボット（クローラ）を国内 5 拠点に設置した(内 2 拠点は 2004 年末に追加)．現時点での収集ロボットの PC 数は合計 70 台（2CPU マシン×10 台，1CPU マシン×60 台）である．これらの PC に，起点となる URL を割り振り，収集するドメインを割り当てて収集を行った．起点となる URL は我々が過去に収集したページから判明している起点をベースに，2003 年 9 月時点でアクセス可能な 600 万の URL とした．起点となった URL のドメインの内訳は表 2の通りである．

表 2 起点 URL のドメインの内訳

ドメイン	起点数
com	3,895,782
org	575,550
edu	146,460
net	576,945
uk	212,799
jp	464,423
us,ca,at	131,788
合計	6,003,747

収集したページは，ベイシスの言語判定[9]により，言語判定が行われる．

5. 解析プラットフォーム

解析を行うプラットフォームについて述べる．

5.1. ハードウェア

使用したマシンは，128 台クラスタと Opteron のワークステーションである．各スペックは以下の通りである．

- 128-ノード COE-クラスタ
  - CPU: Pentium4 2.4GHz
  - Memory: 1GB
  - HDD: 400GB x 2 = 800GB
- ワークステーション
  - CPU: Opteron 2.4GHz x 2
  - Memory 16GB
  - HDD: 300GB x 12(RAID5+spare) x 2 = 4.7TB

5.2. ソフトウェア

- Gfarm（Grid File System） Version 1.2-2
  - グリッド環境を対象とした共有ファイルシステム
  - ペタスケールのストレージ，スケーラビリティのある IO
  - 解析データの保管場所
  - 開発は産総研
- GXP（Grid Explorer） Version 2.01
  - グリッド環境を対象とした分散シェル
  - 多数のノードに一斉にコマンドを投入可能
  - スケジューラーとして利用
  - 開発は東大

5.3. 解析方法

集計処理については，128 台のクラスタを用い，Gfarm 上にデータを保存し，GXP で処理を行った．また，リンク解析はワークステーションで処理を行った．

5.3.1. 集計処理

集計処理の形式としては，1 つのデータを変換する 1 対 1 処理，N 個のデータを 1 つに集約する N 対 1 処理，また，N 個のデータを M 個のキーごとに集約する N 対 M 処理に分類できる．

(a) 1 対 1 処理

処理対象のデータサイズが大きく，計算に時間がかかる場合：この場合，計算をできるだけ均等に行うために，データを分割し，マスターワーカー形式で処理を行った，分割したデータは Gfarm 上に配置し，すべてのノード上から同一パスでアクセスできるようにし，GXP を用いて処理を行った．また計算結果については Gfarm のレプリケーションを用いて複製を作成し，ディスクの障害に備えた．

処理対象のデータサイズが小さい場合や，計算に時間がかからない場合：この場合，ローカルディスク上に中間ファイルを保存し，GXP を用いて並列処理を行った．計算結果については，重要なものはバックアップを別途に保存したが，それ以外は欠損した部分の再計算を行った．

(b) N 対 1 処理

集計するデータが大きい場合：まず各ホスト上で集約を行い，その結果を Gfarm 上に保存し，あるホストにおいて全データの集約を行った．

集計するデータが小さい場合：GXP の集約機能を利用する．特に数え上げる場合が相当する．

(c) N 対 M 処理

M 個のキーを数千個以下にまとめた M’個のバケットとし，処理のフェーズを 3 つに分けて処理をした．

1. Gfarm 上の N 個ファイルに対して，1 対 M’の処理をマスターワーカー形式で処理を行い，中間ファイルをローカルに保存する．
2. ローカルの M’個のバケットに対して，{N/ホスト数 H}対 1 の集約処理を各ホスト上で行い，Gfarm 上に保存する．
3. Gfarm 上の M’個のバケットに対して，{ホスト数 H}対 1 の集約処理をマスターワーカー形式で処理する．

Gfarm Version 1.2-2 では，多数のファイルを作成すると，メタサーバがボトルネックとなり，性能が劣化するため，ファイル数を減らすため，バケットとして処理している．

また，Google の MapReduce[10]と比較すると，1 のフェーズが map,2 のフェーズが Combiner Function に，3 のフェーズが reduce に対応づけることができる．

5.3.2. リンク解析

リンク解析は，ワークステーション上にて boost[11]の graph library を用いて行った．ウェブのリンクに関しては，The Connectivity Server[12]や The WebGraph framework[13] などが存在するが，本研究ではホストレベルでのリンク構造に着目したため，これらを考慮しなかった．また，boost の graph library には，並列化バージョンの, The Parallel Boost Graph Library[14]も存在するが，ページ単位の解析を行うにはよりスケーラビリティな方法を用いなければならない．

6. 統計情報

この節では，4節で述べたデータセットの統計情報を示す．

6.1. TLD(Top Level Domain)の分布

解析対象ページの Top Level Domain（TLD）の分布を示す. 図 2はページ数が多い順に 20 ドメインを選択し，TLD のアルファベット順にグラフにしたものである．com ドメインの起点 URL が多いため，com ドメインが特に多く,またページもかなり多く存在している．図 2中の各 ccTLD(country code Top Level Domain)の国名は，表 3の通りである．

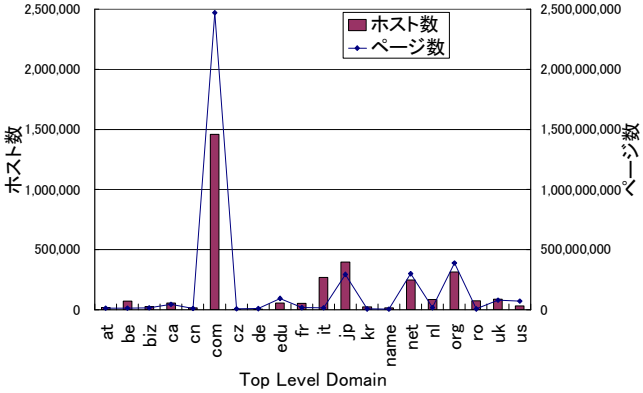

図 2 TLD の分布

表 3 ccTLD の対応表([15]より抜粋)

ccTLD 名	国名
at	Austria
be	Belgium
ca	Canada
cn	China
cz	Czech Republic
de	Germany
fr	France
it	Italy
jp	Japan
kr	Korea, Republic of
nl	Netherlands
ro	Romania
uk	United Kingdom
us	United States

図 3は com ドメインを除いた TLD の分布である. JP ドメインがほかの ccTLD に比べ，多く収集されている．また, it や fr などはホスト数に比べページ数が少ない．逆に edu や net,org などはホスト数に比べページ数が多く，ドメインによってページの偏りがあることがわかる．

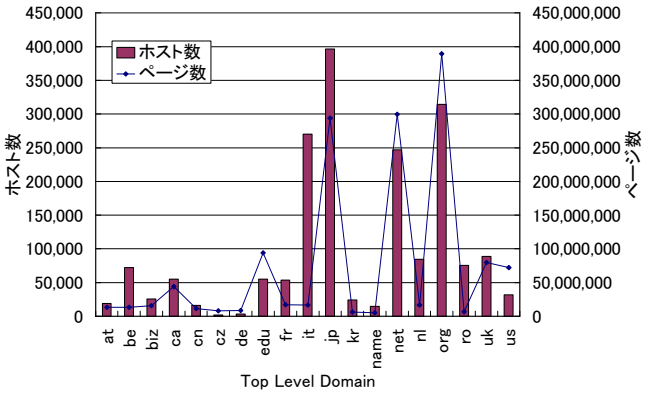

図 3 TLD の分布(com を除く)

また，図 2と図 3のグラフは，ホスト数の軸とページ数の軸の目盛りを 1 ホスト 1000 ページで合わせてあり，折れ線グラフが棒グラフより高い場合は，そのドメインの 1 サーバあたりのウェブページ数が 1000 ページを超えていることを示す．

6.2. 言語の分布

解析対象となるデータセットの言語の分布を示す(図 2)．判定する言語は，**English**, **Japanese**, **Chinese**, **French**, **Korean**, **Spanish**, **German**, **Italian**, **Russian**, **Portuguese**, **Arabic** で，それ以外を **Other** と判定している．**English** が 3 分の 2 のページを占めていた．JP ドメインの起点リストが多いため，日本語が 2 番目に多い．

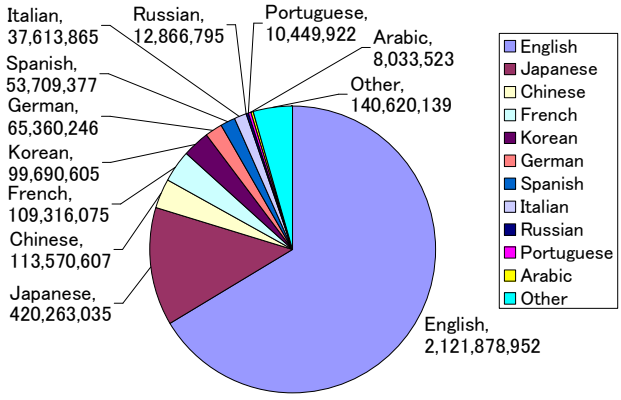

図 4 言語の分布

次に表 4と図 5は言語と TLD の関係を示す表と図である．表 4は各言語の TLD の内訳である．**Japanese** 以外の言語はどれも com ドメインが最も多く，4 割から 7 割近くが com ドメインによって占められている．

表 4 言語ごとの TLD の内訳

Arabic	**com** **69.73%**	net 16.20%	org 10.41%	us 0.87%	ps 0.55%	sa 0.42%	ws 0.35%	bh 0.25%	it 0.17%
German	**com** **40.50%**	at 15.58%	de 11.17%	org 9.68%	net 8.34%	name 6.29%	biz 1.13%	ch 1.01%	uk 0.92%
English	**com** **68.17%**	org 11.77%	net 5.83%	uk 3.61%	edu 3.43%	us 2.13%	ca 1.59%	jp 0.71%	biz 0.66%
Spanish	**com** **67.79%**	org 13.04%	net 6.69%	mx 4.64%	es 1.38%	us 1.16%	edu 0.76%	cl 0.65%	ca 0.45%
French	**com** **58.60%**	fr 11.37%	org 9.09%	ca 7.75%	net 6.18%	be 3.59%	edu 0.95%	jp 0.48%	us 0.42%
Italian	**com** **41.61%**	it 36.74%	org 7.64%	net 7.36%	us 1.28%	edu 0.64%	uk 0.60%	biz 0.55%	jp 0.48%
Japanese	jp 52.79%	**com** **36.97%**	net 7.56%	org 1.55%	tv 0.23%	to 0.22%	st 0.14%	cc 0.14%	bz 0.04%
Korean	**com** **62.19%**	net 18.63%	org 13.16%	kr 5.57%	jp 0.17%	us 0.13%	edu 0.07%	uk 0.01%	biz 0.01%
Portuguese	**com** **51.36%**	org 16.99%	br 11.90%	net 8.34%	pt 2.85%	us 2.63%	edu 1.02%	mx 0.77%	jp 0.61%
Russian	**com** **49.69%**	net 14.70%	org 14.68%	ua 4.71%	ru 2.97%	su 2.81%	kz 2.04%	lt 1.58%	by 1.16%
Chinese	**com** **68.83%**	net 13.96%	cn 9.32%	org 5.39%	jp 0.72%	tw 0.62%	us 0.25%	ca 0.22%	cc 0.22%

図 5は，ページ数の多い 20 個の TLD の言語の分布を示したグラフである．

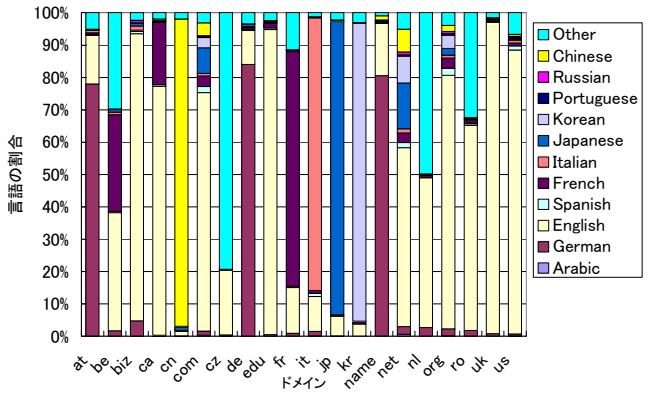

図 5 TLD ごとにおける言語の分布

be ドメイン(**Belgium**)の **Other** は **Belgium** の公用語から推測すると **Dutch** だと考えられる．今，be ドメインの **Other** を **Dutch** と仮定すると，**Dutch**, **France**, **Germany** の割合が公用語の割合と等しいと考えられる([16])．また，cz ドメイン(**Czech Republic**)の **Other** は **Czech** だと推測される([16])．fr ドメイン(**France**)では，公用語は **French** であるが，移民などが多いため **Other** が 1 割ほど占めているのと推測される．name ドメインは，**Germany** が 8 割を占めていた．さらに詳しく調べたところ，**Germany** の一つドメイン仲介取引業者のページが主であった．確認できたホスト数だけで，244 ホスト，3,988,721 ページを持っていた．（表 5は一部抜粋）この業者が持っていると推測されるドメインを除いたところ **Germany** の割合は約 14%まで減り，**English** の割合が約 68%となった．nl ドメイン(**Netherlands**)の **Other** は，**Dutch** だと推測される．ro ドメイン(**Romania**)の **Other** は **Romanian** であると推測される．

表 5 仲介取引業者のドメイン(抜粋)

ホスト名	ページ数	German	English	Spanish	French	Other
algorithm.name	28,732	27,920	810	1	0	1
www.abm.name	22,850	22,192	654	0	1	3
www.adolf-hitler.name	15,673	15,308	365	0	0	0
www.adolf.name	15,668	14,944	723	1	0	0
www.affe.name	16,886	16,700	186	0	0	0
www.afrika.name	24,658	24,350	274	0	0	34
www.alberteinstein.name	15,369	14,928	441	0	0	0
www.algorithm.name	20,438	19,664	773	0	0	1
www.almanach.name	17,108	17,014	88	0	0	6
www.alternativmedizin.name	11,469	11,433	36	0	0	0

6.3. 静的・動的ページ

収集対象のページは，リンクが存在する場合，ホスト上にあるファイルである静的ページではなく，CGI などによって生成される動的ページも収集される．具体的には，

http://www.infoseek.co.jp/Keyword?pg=ranking_news_if.html&svx=120&sv=KW#sports

の‘?’を含む URL のページも収集されている．解析対象のデータセットは，ページ数は 3,935,592,289 ページでホスト数は 3,652,232 ホストなので，1 ホストあたり約 1077.6 ページあることになる．

しかし，これまでの調査において，[1]の 1998 年では 1 ホストあたり 190 ページ，[2]の 1999 年の調査では 186 ページ，また[3]の 2004 年の調査では 202 ページとなっており，1 ホストあたり 200 ページ前後と推定され，解析対象の平均ページ数と大きく異なる．

この要因としては，CGI 等によって生成される動的ページの増加が考えられる．以下では，動的ページの割合を調査するために URL 中に'?'が含まれている URL を動的ページ，'?'を含まない URL を静的ページと仮定して調査を行った．

調査の結果，'?'を含むページは 2,294,025,470 ページで，逆に含まないページは 1,641,566,819 ページで，動的ページと静的ページの割合は約 6 対 4 となった．よって，'?'を含まない URL を静的ページと仮定した場合，一ホストあたり約 450 ページとなる．

さらに，動的ページと静的ページの特徴を調べるため，ディレクトリの階層の深さごとにページの数が異なるか調査を行った．（図 6）

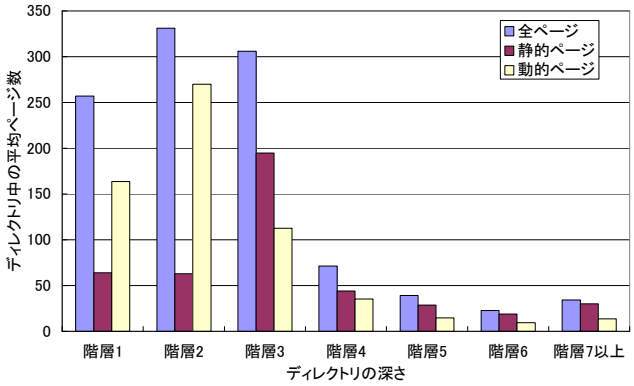

図 6 ディレクトリの深さとページ数

図 6の階層 1 とは，

http://www.hoge.com/index.html
http://www.hoge.com/index2.html
http://www.hoge.com/index.php?abc=hoge

http://www.hoge.com/index.php?abc=fuga

などの Web サーバのルートのディレクトリにある平均のページ数で，階層 2 とは，

http://www.hoge.com/fuga/index.html
http://www.hoge.com/fuga/a.html

の平均ページ数で，階層 3，階層 4 とそれぞれのディレクトリの深さの平均ページ数である．

この結果動的ページは，階層 2 まで増加し，階層 3 から減少しており，動的ページは深い階層には比較的少ないことがわかる．また，静的ページでは階層 3 が突出して多い．これについて調査したところ，あるポルノサイトやドメイン仲介業者が多数のホスト，多数のページを持っていることがわかった．

これを調べるに当たり，次の手順で調査を行った．

1. ホストを，1 ホストあたりのページ数ごとに適当な区間で分割した
2. 分割したホストごとに，階層ごとのページ数の類似度でクラスタリングを行った．
3. 各クラスタでホスト数や総ページ数が多いクラスタからサンプリングを行いどのようなホストが集まったか調査した．

この調査の結果，同様なホストが集まっていたクラスタのメディアンを表 6に示す．比較のために静的ページも示す．

表 6 階層の類似度によるクラスタ

	ページ数	階層 1	階層 2	階層 3	階層 4	階層 5
静的ページ	442.82	63.96	62.81	194.85	43.93	28.42
主にポルノ	49,847.74	94.57	522.84	48,539.62	395.72	141.29
主にポルノ	104,577.46	20.40	313.07	103,987.70	206.79	12.28
主にポルノ	171,205.16	31.98	182.11	170,333.76	500.90	132.60
主に業者	93,324.76	92,259.81	703.88	187.12	52.80	42.84
主に業者	181,345.32	180,758.71	257.93	226.15	101.75	0.71
主に業者	342,240.34	340,896.38	1,182.98	146.06	10.82	3.97

また，表 6のクラスタに含まれるホストと’?’を含む URL を除いた場合のディレクトリの深さごとの平均ページ数を図 7に示す． このときのページ数は，993,141,984 ページで，ホスト数は 3,616,797 ホストとなり，ホストあたりの平均ページ数は 274.59 となった．

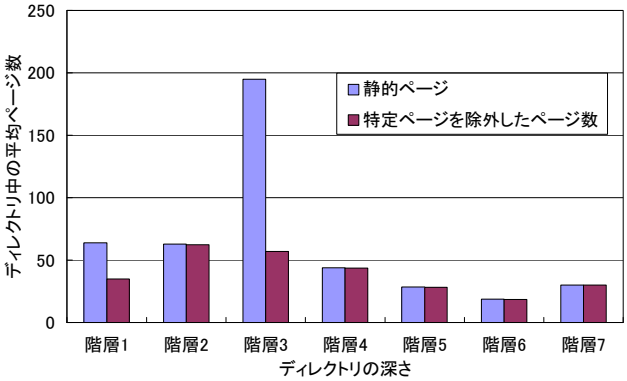

図 7 特定のページを除外した場合

7. 強連結成分の解析

強連結成分の解析は，グラフ理論の強連結成分抽出を TLD ごと，言語ごとに行う．また全体の概要を把握するために，解析対象のグラフはページ単位ではなく，ホスト単位とした．つまり，ホストが持つすべてのページを一つの頂点とみなし，他のホストへのリンクすべてをその頂点からの辺としたホストグラフを解析対象とした

また，解析に使用したデータセットは，総ページ数 3,208,139,905 ページ，ページ間の総リンク数は 93,397,065,743 リンクである．ページ間の総リンク数の内，ホスト内に閉じている**内部リンク数**は，77,971,241,488 で，ホスト外へリンクしている**外部リンク**は，15,425,824,255 である．ホスト数は 1,719,134 ホスト，ホスト間のリンク数は 91,084,879 である．

7.1. TLD ごとのホストグラフ

表 7に各 TLD に分割したホストグラフの統計を示す．

表 7 TLD ごとのホストグラフ

TLD	ホスト数	ページ数	内部リンク	外部リンク	同一ドメインへのリンク
com	1.31M	2.76G	66G	14G	14G
de	3.02K	8.43M	322M	2.07M	77K
edu	80	600K	11M	242K	161
fr	586	563K	18M	359K	894
it	593	1.57M	46M	218K	19K
jp	380K	369M	9.67G	1.03G	950M
kr	209	2.40M	62M	25K	590
net	5.82K	15M	422M	25M	5.40M
org	3.74K	12M	317M	7.83M	1.52M
ru	141	450K	26M	265K	301

また，TLD ごとのリンク先 TLD の分布を示す(表 9)．JP ドメイン以外のドメインでは，COM ドメインへのリンクが多い．これは，JP ドメインを初期に集中して収集した影響だと考えられる．

表 8 TLD ごとのリンク先 TLD の分布

TLD	外部リンク	com	de	edu	fr	it	jp	kr	net	org	ru	oth
com	14G	98.80%	0.01%	0.00%	0.01%	0.00%	0.89%	0.00%	0.15%	0.08%	0.00%	0.05%
de	2.07M	84.41%	3.73%	0.03%	0.10%	0.14%	3.65%	0.03%	0.06%	1.83%	0.02%	6.00%
edu	242K	99.01%	0.00%	0.07%	0.00%	0.00%	0.31%	0.00%	0.08%	0.40%	0.00%	0.13%
fr	359K	99.09%	0.00%	0.00%	0.25%	0.00%	0.16%	0.00%	0.21%	0.09%	0.00%	0.20%
it	218K	88.36%	0.12%	0.04%	0.02%	8.74%	1.00%	0.00%	0.04%	0.82%	0.08%	0.79%
jp	1.03G	7.34%	0.01%	0.00%	0.00%	0.00%	91.88%	0.00%	0.42%	0.08%	0.00%	0.26%
kr	25K	93.41%	0.00%	0.00%	0.00%	0.00%	1.66%	2.35%	0.06%	2.48%	0.00%	0.04%
net	25M	33.74%	0.03%	0.00%	0.01%	0.01%	41.00%	0.00%	21.71%	1.18%	0.00%	2.32%
org	7.83M	47.46%	0.33%	0.00%	0.00%	0.01%	29.53%	0.02%	1.95%	19.47%	0.00%	1.23%
ru	265K	99.66%	0.00%	0.02%	0.00%	0.00%	0.05%	0.00%	0.00%	0.14%	0.11%	0.02%
oth	14M	45.15%	0.62%	0.01%	0.03%	0.01%	23.50%	0.00%	0.59%	0.89%	0.01%	29.20%

7.2. 言語によるホストの分類

言語別の解析をホスト単位で行うため，ホストを言語ごとに分類する必要がある．分類するにあたりホストで最も使用されている言語を元に分類した．ただし，複数の言語を使用しているホストは解析対象外とした．

そのため，まずホストを分類するにあたり，複数の言語を使用するホストを除外した．図 8は最も使用されている言語で何割のページが占有しているかを横軸に，そのホストの割合を縦軸とした図である．

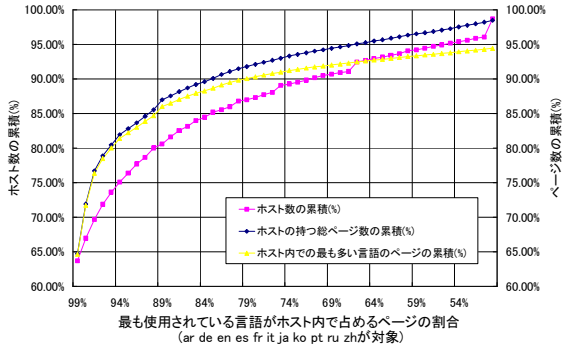

図 8 単一言語の占有率

図 8によると，ホストが保有する 7 割以上のページが単一の言語のみを使用しているホスト数は 9 割を占めている．よって，ホストの分類はホストの保有するページの 7 割が単一の言語であるホストを抽出し分類した．(図 9)

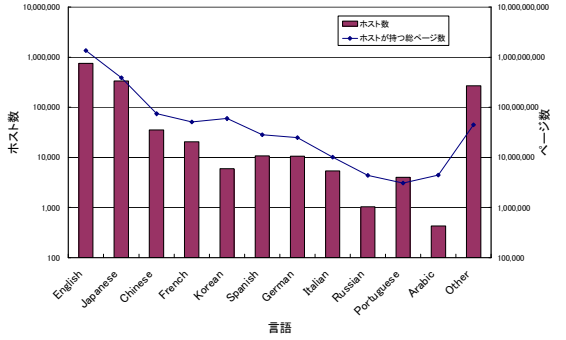

図 9 言語別ホストのホスト数とページ数

また，分類したホストが適切に分類されているか調べるため，分類したホストが保有するページの言語の割合を調査した．表 9は縦に分類，横に言語割合を示した表で，ホストごとの言語の割合を求め，分類内で平均した値を示したものである．どの分類も 9 割以上分類した言語によって，占められていた．

表 9 各分類における言語の分布

	Arabic	German	English	Spanish	French	Italian	Japanese	Korean	Portuguese	Russian	Chinese	Other
Arabic	95.50%	0.01%	2.82%	0.08%	0.10%	0.08%	0%	0%	0.02%	0%	0%	1.34%
German	0%	95.26%	2.45%	0.05%	0.17%	0.41%	0.02%	0%	0.02%	0%	0%	1.56%
English	0%	0.22%	97.35%	0.11%	0.30%	0.51%	0.07%	0%	0.03%	0%	0.02%	1.33%
Spanish	0%	0.08%	2.39%	94.87%	0.18%	0.47%	0.01%	0%	0.43%	0%	0%	1.52%
French	0%	0.13%	2.47%	0.17%	95.07%	0.46%	0.02%	0%	0.03%	0%	0.01%	1.59%
Italian	0%	0.22%	3.16%	0.12%	0.20%	93.52%	1.38%	0%	0.04%	0%	0%	1.30%
Japanese	0%	0.04%	1.95%	0.02%	0.08%	0.14%	95.16%	0%	0%	0%	0.03%	2.53%
Korean	0%	0.07%	1.70%	0.01%	0.19%	0.15%	0.06%	95.44%	0.01%	0%	0.09%	2.22%
Portuguese	0%	0.04%	1.56%	0.23%	0.08%	0.19%	0%	0%	97.22%	0%	0%	0.63%
Russian	0.01%	0.07%	3.10%	0.02%	0.09%	0.52%	0.03%	0%	0%	94.55%	0%	1.55%
Chinese	0%	0.03%	1.80%	0%	0.15%	0.35%	0.04%	0.01%	0%	0%	96.52%	1.05%

7.3. 言語ごとのホストグラフ

表 10に7.2で行ったホスト分類ごとに分割した各言語のホストグラフの統計を示す．

表 10 言語ごとのホストグラフ

言語	ホスト数	ページ数	内部リンク	外部リンク	同言語へのリンク
Arabic	422	4.96M	99M	8.61M	5.90M
Chinese	35K	86M	2.60G	253M	196M
English	756K	1.57G	48G	4.03G	3.81G
French	20K	59M	2.26G	79M	55M
German	11K	27M	860M	17M	2.42M
Italian	5.41K	11M	440M	14M	4.41M
Japanese	336K	519M	13G	2.50G	1.36G
Korea	5.97K	66M	1.55G	47M	40M
Portuguese	3.75K	3.14M	66M	1.85M	651K
Russian	1.02K	4.75M	276M	3.14M	1.54M
Spanish	11K	31M	1.20G	36M	23M
Other	534K	825M	8.19G	8.44G	2.48G

また，言語ごとのリンク先言語の割合を示す(表 11).

表 11 言語ごとのリンク先言語の割合

言語	外部リンク	Ar	Ch	En	Fr	Ge	It	Ja	Ko	Po	Ru	Sp	Ot
Arabic	8.61M	69%	0%	11%	0%	0%	0%	0%	0%	0%	0%	0%	21%
Chinese	253M	0%	78%	8%	0%	0%	0%	0%	0%	0%	0%	0%	14%
English	4.03G	0%	0%	95%	0%	0%	0%	1%	0%	0%	0%	0%	4%
French	79M	0%	0%	19%	69%	0%	0%	0%	0%	0%	0%	0%	11%
German	17M	0%	0%	65%	2%	14%	2%	2%	1%	0%	0%	2%	13%
Italian	14M	0%	0%	39%	1%	1%	33%	3%	0%	0%	0%	1%	22%
Japanese	2.50G	0%	0%	2%	0%	0%	0%	54%	0%	0%	0%	0%	44%
Korea	47M	0%	0%	6%	0%	0%	0%	0%	85%	0%	0%	0%	9%
Portuguese	1.85M	0%	1%	42%	1%	0%	0%	1%	1%	35%	0%	4%	16%
Russian	3.14M	0%	0%	36%	0%	0%	0%	0%	0%	0%	49%	0%	15%
Spanish	36M	0%	0%	15%	0%	0%	0%	1%	0%	1%	0%	65%	18%
Other	8.44G	0%	0%	11%	0%	0%	0%	60%	0%	0%	0%	0%	29%

7.4. 強連結成分の解析結果

表 12は，ホストレベルでの強連結成分の解析結果を示す．

表 12 ホストレベルでの蝶ネクタイ構造

	CORE	IN	OUT	Other	Total
ホスト数	624,173	147,794	621,788	325,379	1,719,134
ホストの割合	36.30%	8.60%	36.20%	18.90%	
ホストが保有するページ数	2,102,971,321	633,530,035	346,251,616	125,386,933	3,208,139,905
ページの割合	65.60%	19.70%	10.80%	3.90%	
ホストあたりのページ数	3,369.20	4,286.60	556.9	385.4	1,866.10

表 12から推測すると，Broder らの調査した 1999 年のウェブよりも CORE が大きいと予想され，また，Lie らは，China Web の CORE の巨大化は，中国特有の現象と主張していたが，全世界的にも CORE は巨大化していると予想される．

7.5. TLD ごとの強連結成分

TLD ごとにウェブを分割して,ホストレベルで強連結成分の解析を行い，各成分のページの割合を表 13 に示す．

表 13 TLD ごとの強連結成分

TLD	ページ数	同一ドメインへのリンク	SCC	IN	OUT	Other
com	2.76G	14G	53.65%	19.73%	22.25%	4.37%
de	8.43M	77K	0.25%	0.05%	78.36%	21.34%
edu	600K	161	0.05%	0.00%	14.44%	85.51%
fr	563K	894	0.01%	0.02%	25.33%	74.63%
it	1.57M	19K	0.11%	0.04%	0.04%	99.81%
jp	369M	950M	26.46%	1.77%	71.32%	0.46%
kr	2.40M	590	0.00%	0.00%	1.09%	98.91%
net	15M	5.40M	0.52%	0.17%	35.42%	63.89%
org	12M	1.52M	0.61%	0.38%	64.25%	34.76%
ru	450K	301	0.77%	0.05%	0.49%	98.70%

その結果，com ドメインと jp ドメイン以外は，巨大な強連結が存在せず，jp ドメインでも 26%と Broder らの調査の調査よりも小さくなった．ここから推測されることは，ウェブはドメインごとには分かれていないことが予想される．

7.6. 言語ごとの強連結成分

7.2で行ったホスト分類ごとに強連結成分を解析した結果が，表 14で，各成分のページの割合を示した．

表 14 言語ごとの強連結成分

言語	ページ数	同言語へのリンク	CORE	IN	OUT	Other
Arabic	4.96M	5.90M	61.43%	10.20%	18.59%	9.78%
Chinese	86M	196M	76.88%	9.98%	10.57%	2.57%
English	1.57G	3.81G	66.90%	9.04%	16.44%	7.62%
French	59M	55M	61.85%	9.23%	20.65%	8.27%
German	27M	2.42M	26.61%	8.16%	42.18%	23.05%
Italian	11M	4.41M	23.67%	17.10%	29.54%	29.69%
Japanese	519M	1.36G	71.05%	25.85%	2.54%	0.56%
Korea	66M	40M	54.32%	17.07%	19.36%	9.25%
Portuguese	3.14M	651K	26.60%	4.94%	42.18%	26.28%
Russian	4.75M	1.54M	35.76%	18.20%	18.35%	27.69%
Spanish	31M	23M	64.93%	5.30%	23.60%	6.16%
Other	825M	2.48G	7.24%	1.98%	9.32%	81.47%

Chinese の構成比を見ると，Lie らの調査した China Web と似たような比になっており，**Chinese** は他の言語と比べても CORE の比が最も大きくなっている．ただし，日本語のページも同じような構成比となっている．

8. おわりに

本稿では，2006 年 2 月までに収集をした 120 億ページを元に推定すると 2006 年 2 月時点で 350 億ページが存在するという結果を得た．

また，約 30 億のウェブを対象に，ホストレベルでの強連結成分を行った．また，ホストを **Top Level Domain** や主要言語別に分類し，それぞれの強連結成分を比較した．

その結果，1999 年のウェブに比べると CORE が巨大化していることが判明した．**Top Level Domain** ごと解析よると，ウェブは **Top Level Domain** ごとには分かれていないことが判明した．また，主要言語別の解析によれば，中国語や日本語ではより **CORE** が巨大化する傾向にあることが判明した．

今後の課題としては，解析の対象が収集したページの一部であったが，収集した全ページを対象に調査を行う．また，強連結成分の解析は，ホストレベルの解析であったが，ページレベルでの解析を行い，より詳細調査を行う．

謝　辞

本研究の一部は，文科省 21 世紀 COE「プロダクティブ ICT アカデミア」及び科学技術振興費「e-Society」プロジェクトによるものである．

富士通株式会社及び FFC システムの皆様に深く感謝いたします．

文　献

[1] S.Lawrence, C.L.Giles:"Searching the World Wide Web", Science, Vol.280, No.5360, pp.98-100　1998)

[2] S.Lawrence, C.L.Giles:"Accessibility of Information on the Web", Nature, Vol.400, pp.107-109　1999)

[3] 総務省情報通信政策研究所:WWW コンテンツ統計調査報告書, http://www.soumu.go.jp/iicp/chousakenkyu/seika/houkoku.html　2004.7)

[4] Netcraft Home Page, http://www.netcraft.co.uk/

[5] A. Broder, R. Kumar, F. Maghoul, P. Raghavan R. State, A. Tomkins, and J. Wiener. Graph structure in the web, Proc. 9th World Wide Web Conf. 2000.5)

[6] P. Boldi, B. Codenotti, M. Santini, and S. Vigna. Structural properties of the African web. 2002.

[7] G. Lie, Y. Yu, J. Han, G. Xue: China web graph measurements and evolution, Proc. Asia Pacific Web Conf., LNCS, Vol.3399,pp668-679 2005.3)

[8] e-Society プロジェクト http://www.yama.info.waseda.ac.jp/~yamana/es/

[9] Basis Technology Rosette 言語判別システム http://www.basistech.co.jp/language-identification/

[10] Dean, J. and Ghemawat, S. MapReduce: Simplified Data Processing on Large Clusters in OSDI'04: 6th Symp, 2004.

[11] BOOST. http://www.boost.org

[12] K. Bharat, A. Broder, M. Henzinger, P. Kumar, and S. Venkatasubramanian. The connectivity server: fast access to linkage information on the web, Proc. 7th WWW, 1998.

[13] Paolo Boldi and Sebastiano Vigna. The WebGraph framework I: Compression techniques. In Proc. of the Thirteenth International World Wide Web Conference, pages 595-601, Manhattan, USA, 2004. ACM Press.

[14] D. Gregor, N. Edmonds, B. Barrett, and A. Lumsdaine. The Parallel Boost Graph Library. http://www.osl.iu.edu/research/pbgl, 2005.

[15] IANA : Root-Zone Whois Index by TLD Code, http://www.iana.org/cctld/cctld-whois.htm

[16] Wikipedia: フリー百科事典『ウィキペディア（Wikipedia）』, http://ja.wikipedia.org/