World Wide Web の測定：Web ページ推定手法の比較

池内淳（大東文化大学）
安形輝（亜細亜大学）

Ⅰ．はじめに

これまで、インターネット上にある Web ページ数を測定するための手法が二、三考案されてきた。それらの結果、Web が「指数関数的」と形容されるような、極めて急速なペースで成長し続けていることや、サーチエンジンの収載能力が、その成長に追いついていないこと等が明らかにされてきた。

しかしながら、それらの既往調査は、異なった方法論を用いて、散発的に行われてきたに過ぎず、調査結果間の相互比較が必ずしも円滑に行われるという訳ではない。また、それぞれの手法には、いずれも、方法論上の潜在的な欠陥の存在することが既に知られているが、それが測定の際にどのような影響を及ぼし、実数と推定値にどの程度の誤差を生じさせることになるのかを明らかにする手だてはない。同様に、いずれの方法論に則れば、最も精度の高い測定が行われるかといった論点についても、これまで顧みられることはなかったと言える。

そこで、本研究では、同時期に、二つの測定手法を用いて Web ページ数を推定することにより、各々の調査結果によって導き出される数値に、どの程度の差異が存在するのかを実証的に明らかにするとともに、各々の方法論がどのような特性を持っているのかについて考察を加える。

Ⅱ．既往調査

原田 [1)]は、Web ページ数を推定するための手法を、以下の三つのアプローチに分類している。

（1）複数のサーチエンジンを用いる方法
（2）WWW サーバの無作為抽出による方法
（3）調査ロボットを用いた方法

本研究では、このうち、(1)と(2)の方法論について取り上げる。

A. 複数のサーチエンジンを用いる方法

ここでは、一対（もしくは多対）のサーチエンジン間の出力した URL（群）の一致数から、Web ページ数が推定される。これは、生物学分野において「捕獲―再捕獲法」などと呼ばれ、広く用いられている手法を応用したものである。このアプローチに基づいた事例としては、Bharat & Broder[2), 3)]や Lawrence & Giles[4)]による調査の存在するほか、我が国の Web ページのみを対象とした調査もいくつか存在する。

インターネット上に存在する Web ページ数の推定値を N とすると、N は 1 式によって求めることができる。

$$N = \frac{n_1 n_2}{m} \qquad (1)$$

ここで、n_1、n_2 はサーチエンジンの収集ページ数、m は n_1 と n_2 中の一致したページ数である。

但し、1 式が妥当であるのは、n_1 及び n_2 が母集団から無作為に抽出された場合のみである。言うまでもなく、ロボットにおるクローリングの場合、被リンク数の多寡によって、サーチエンジンに収集される確率は大きく変化すると考えられるし、実際に、頻繁に参照されているサーバについては、
より深いレベルまでクローリングを行うといった方針を採用しているロボットも存在する。したがって、ここでの推定値は、実際の Web ページ数を過小に評価してしまう虞がある。

B. WWW サーバの無作為抽出による方法

「WWW サーバの無作為抽出による方法」を用いた事例としては、Lawrence & Giles[5]による事例が存在する。

ここでは、まず、現行の標準規格である IPv4 で可能な 32 ビット（＝4,294,967,296）の IP アドレスに対して、ランダムなアクセスを繰り返し、Web ポート（80 番）の開いているサーバ数をチェックすることによって、全体に占める WWW ホストの比率を確認する。さらに、1 サーバ当たりの平均 Web ページ数を測定し、両者を掛け合わせることによって、Web ページ全体を推定している。

この手法は、Web サーバへのアクセスの段階では、三つのアプローチのうち最も偏りのないサンプリングが行えることに疑いはないが、サーバ単位の平均ページ数が正確に測定できるかどうかという難点が存在する。また、仮に、正確に測定できていたとしても、ごく一部の極めて多くのページを保持しているサイト（プロバイダ等）を含むか否かによって、この値が大きく変化してしまうこととなる。

C. 既往調査による推定値の変遷

以上のような方法論に則って、これまでに行われた既往調査による結果を表1に示す（→表1）。

これに従えば、1997 年 6 月から 1999 年 2 月までの Web ページの平均増加率増加数は、一ヶ月、約 3300 万ページであるということになる。また、Web 全体に占める最大のサーチエンジンの索引率についても、1997 年当初は最大で 50％程度もあったものが、1999 年の調査では、16％にまで減少しており、Web の急速な成長にサーチエンジンが追ついていないことなどが描写されている。

表 1. 既往調査による Web ページ数

	調査時期	調査方法	Web ページ数
Bharat ら[2]	Jun. 1997	サーチエンジン	125,000,000
Bharat ら[2]	Nov. 1997	サーチエンジン	200,000,000
Lawrence ら[4]	Dec. 1997	サーチエンジン	320,000,000
Bharat ら[3]	Mar. 1998	サーチエンジン	250,000,000
Lawrence ら[5]	Feb. 1999	Web サーバ	800,000,000

Ⅲ. 方法論と結果

本稿では、二つの方法論のうち「複数のサーチエンジンを用いる方法」に基づいた調査の結果について述べる。

調査は以下の手順にしたがって行われた。

① サーチエンジンの選定
② 標本 URL の抽出
③ URL の正規化と重複除去
④ 一致数の測定
⑤ サーチエンジンの規模の測定
⑥ Web ページ数の推定

まず、調査のために使用するサーチエンジンについては、規模・安定性・知名度などの要素を勘案して、以下の四つを選定した。

① AlltheWeb: http://www.alltheweb.com/
② AltaVista: http://www.altavista.com/
③ Google: http://www.google.com/
④ HotBot: http://www.hotbot.lycos.com/

このほかにも Northern Light, Lycos, Excite, WiseNut など調査対象の候補となるサーチエンジンは複数あったが、それぞれ、なんらかの理由により除外した。

次に、標本 URL の抽出については、WordNet のシソーラス（約 19 万語）から単語（約 8 万語）を切り出し、それらの中から無作為に 2 語を組み合わせて AND 検索を行った。

ここでは、各サーチエンジンに対して 1 万回ずつ検索を繰り返し、四つ全てのサーチエンジンで 1 件以上ヒットし、かつまた、各サーチエンジンの最大表示件数（ほとんどが 1,000 件）以内に収まるもののみを対象とした。

この作業は2001年8月30日～31日にかけておこなわれ、その結果、4,224件(42.2%)が採用された。この段階で収集された延べURL数は1,192,728であった。

次に、以下の手順にしたがって、URLの正規化を行うとともに、重複除去を行った。

① URLを小文字に統一
② 16進数文字の変換
③ IPをドメインネームに変換
④ ポート番号の除去
⑤ デフォルト・ページの除去
⑥ バイナリ・ファイルの除去
⑦ 各サーチエンジン内での重複除去

ちなみに、⑤のデフォルト・ページとは、ここでは、index.html, index.htm, default.aspを指す。

以上の結果、四つ全てのサーチエンジンの重複除去後のURLの合計は930,598(78%)となった。

次に、サーチエンジン間の出力URLの一致数を求めた。標本全体のユニークURLの数は603,934(50.6%)であった。ここで、ユニークURLのうち、単一のサーチエンジンのみしか出力しなかったもの、あるいは、複数のサーチエンジンが重複して出力したものの比率を示したものが表2である。(→表2)

四つ全てにおいて等しく検索されたものは全体の4.5%、単一のサーエンジンのみしか出力しなかったものが64.4%にのぼった。

さらに、その64.4%(388,842)の内訳を表3に示した(→表3)。

表2. ユニークURLに占めるサーチエンジン間の重複数

overlap by 4	26,902(4.5%)
By 3	57,770(9.6%)
By 2	130,420(21.6%)
By 1	388,842(64.4%)
unique URL	603,934(100%)

表3. 単一のサーチエンジンによる出力の内訳

AlltheWeb	114,944(29.6%)
AltaVista	47,857(12.3%)
Google	152,423(39.2%)
Hotbot	73,618(18.9%)
total	388,842(100%)

四つのサーチエンジン間の出力URLの一致率は、表4の通りである(→表4)。概ね20%～50%の間に落ち着いていることが看て取れる。

次に、サーチエンジンの絶対数を測定した。多くのサーチエンジンが公表している収集ページ数は、かならずしも正確であるとは言えないため、ここでは、独自に調査を行った。

URLからの検索が可能なAltaVistaとAlltheWebを用いて、URL中に「http」を含むページを全て検索したところ、前者は525,199,572、後者は620,078,439という値が得られた。さらに、AltaVistaのドメイン検索機能を用いて250のトップレベル・ドメインごとに検索を行い、その合計数を求めたところ、577,089,095となった。

以上の数値と、今回の調査から得られた各サーチエンジンの出力数の比率とを掛け合わせ

表4. サーチエンジン間のURLの一致率

	Alltheweb	AltaVista	Google	Hotbot
Alltheweb	—	21.8%	47.9%	26.5%
AltaVista	45.0%	—	52.4%	31.6%
Google	41.5%	22.0%	—	22.7%
Hotbot	42.5%	24.5%	41.9%	—

表 5. 推定された Web ページ数

	Alltheweb	AltaVista	Google
AltaVista	2,649,690,773		
Google	2,870,789,450	2,622,132,982	
Hotbot	2,804,882,823	2,356,332,563	3,281,708,672

ることによって、サーチエンジンのデータベースの規模（＝検索可能ページ数）を推定した。基準値としては、最も大きい AltaVista の 577,089,095 を採用することとした。

以上のデータを用いて、2001 年 8 月 31 日時点の Web ページ数を求めた。各サーチエンジンの対ごとの Web ページの推定値を表 5 に示した（→表 5）。

最も少ないのは、AltaVista と Google による約 26 億ページであり、最も多かったのは Google と HotBot による 32 億ページであった。この方法論による推定値は Web ページ数を過小に評価してしまう可能性があることから、ここでは、既往調査と同様、最大の推定値を採用することとした。

その結果、この中で最大のサーチエンジンである Google は Web 全体の 41.9%を網羅しており、既往調査の結果に反して、Web の成長に十分追随している。また、四つのサーチエンジンを併用いれば、全体の 76.9 % を検索することが可能であることが分かる。

【注・引用文献】

1)原田昌紀．“WWW ロボットとサーチエンジンのスケーラビリティ”．bit．Vol.31，No.12，p.22-28(1999)

2)Bharat，K.，Broder，A.“Estimting the Relative Size and Overlap of Public Web Search Engines”. WWW7 in April 1998.

3)Bharat，K，Broder，A.“Measuring the Web”. http://www.research.digital.com/SRC/whatsnew/sem.html

4)Lawrence，s.，Giles，C. L.“Searching the World Wide Web”. Science. Vol.280，No.5360，p.98-100(1998)

5)Lawrence，S.，Giles，C. L.“Accessibility of Information on the Web”. Nature. Vol.400，p.107-109(1999)