検索エンジンを利用した日本語 Web ページ数の統計的推定

来住伸子[†] 大森貴博[††]
水谷正大[††] 小川貴英[†]

Web ページ数の増大にともない，Web に関する統計調査方法の研究が重要になってきた．この研究では，Lawrence らが 1998 年に提案した検索エンジンの検索結果の重複を利用した推定方法を使い，日本語の Web ページ数の統計的推定を行った．その結果，2000 年 10 月に少なくとも約 2 億 5,600 万個の日本語の Web ページが存在していたと推定できた．また，この推定方法が仮定しているモデルに関する検討と継続的な利用に関する問題点などについて考察を行った．

Statistical Analysis of Japanese Web Pages Using Search Engine Coverage

Nobuko Kishi,[†] Takahiro Ohmori,[††] Masahiro Mizutani[††] and Takahide Ogawa[†]

The rapid growth of the Web has made it impossible to learn directly various properties of the entire Web. Lawrence, et al. proposed a method for estimating the number of Web pages using search engine coverage overlap. We have applied their method to Japanese Web pages and have obtained an estimate of 256 million pages as a lower bound on the size of Japanese indexable Web in October, 2000. We also discuss the limits and possibilities of search engine coverage overlap methods for measuring the Web in the future.

1. はじめに

ここ数年のインターネット普及の伸びは目覚ましいものがあり，World Wide Web（Web）ページ数は急増している．しかし，どのくらいの数の Web ページが実際に存在するか，どのくらいの数の Web ページを検索エンジンで実際に検索できるのか，についての正確なデータが現状ではほとんどない．そのため，Web 技術の基礎研究として，Web に関する統計調査方法の研究は非常に重要になってきた．

1997 年に Lawrence らが英語の検索エンジンを使った英語の Web ページ数の推定を行った[1)]．彼らは，英語の Web ページ数は 1997 年には最低約 3 億 2,000 万ページあると推定し，当時の最大の検索エンジンでも約 34%のページしか検索できないことを指摘した．そこで，日本語の検索エンジンを用いて同じ方法を使うことにより，日本語 Web ページ数の推定を行うことにした．これにより，日本語 Web ページ数の英語の Web ページ数との比較や，Lawrence らの方法の仮定するモデルや推定値の統計的信頼性について詳細な検討をすることが可能になると考えたためである．調査の結果，2000 年 10 月に日本語の Web ページは最低約 2 億 5,600 万個あることが推定できた．また，我々が以前に調査した値も含めると，日本語の Web ページ数は 1999 年から 2000 年にかけて，通信白書で示されている増加より急激に増加していることが分かった．一方，検索エンジンを利用した推定方法はクエリの選定などにかなりの手間がかかり，長期間にわたる調査には適用できない可能性があることが分かった．

本稿では，2 章で従来の Web の調査方法を紹介し，3 章では，Web ページ数の推定を中心に，Lawrence らのモデルに基づく統計的推定の詳細な説明を行う．つづいて，4 章では我々が実際に使用した実験方法を述べ，5 章でその実験結果を紹介する．5 章では日本語 Web ページの数の推定値のほかに，無効 URL と無効ページ存在率も紹介する．6 章では実験結果を考察し，仮定するモデルの妥当性やこの推定方法の限界について考察する．

† 津田塾大学
Tsuda College
†† 東京情報大学
Tokyo University of Information Sciences

2. 従来の研究と実験の動機

1989年にWebの構想が提案され，1991年に最初のWebシステムが提供された．1993年初めに，Mosaicが提供されるとともに，急速にWebサーバ数が増加した[2]．1993年から2年間ほどはWebがまだ普及していなかったため，検索エンジンが実際に集めたWebページ数が世界のWebページ数と考えられた．たとえば，1995年11月にOpenTextは11,366,121個のURLについて数種の統計量を報告している[3]．しかし，Webページ数の増加とともに，単独の検索エンジンではすべてのWebページを収集できなくなった．

そこで，Bharatらは，複数の検索エンジンを利用し，異なる検索エンジンの検索結果として得られたURL群の中で重複するURLを利用して，Webページの調査を行うことを提案した[4]．しかし，Bharatらの方法は，重複するURL集合の計算方法などに問題があり，その問題点を改善した方法を提案したのがLawrenceらである[1]．Lawrenceらは，米国NECの研究所所員が検索エンジンを利用した履歴から302個のクエリを選び，それらをAltaVista，Excite，HotBot，Infoseek，Lycos，Northern Lightの6検索サービスに与えた．検索結果として得られたWebページ集合から，各検索サービス間の重複集合を計算し，検索可能な公開されたWebページ数が1997年12月には最低3億2,000万ページであったと推定した．さらに，1999年2月にも同様の調査を行い最低3億3,500万ページと推定している[5]．また，1999年2月の調査では，IPアドレスのランダム抽出によるWebページ数の推定も行っており，英語のページに限定しない，日本語を含むすべての言語のWebページ数は8億個と推定している．

日本では，平成12年版通信白書に1998年2月に1,020万ページ，1998年8月に1,790万ページ，1999年2月に2,950万ページ，1999年8月に3,850万ページ，という数値が記載されている[6]．これらの数値は，郵政省郵政研究所による調査[7),8)]に基づいており，この調査は，実際に収集したWebページ数から線形予測した値を採用している．一方，検索エンジンgooは1998年6月26日に1,700万ページ，1999年11月11日に3,500万ページ[9),10)]，検索エンジンLycos日本語版は1999年5月17日に3,000万ページ[11)]を収集したことを公表している．これらの数をそのまま解釈すると，日本のWebページの7割以上が単独の検索エンジンで検索できていることになるが，これは，多くのユーザの実感に合わない．したがって，通信白書で示された数値は，日本語Webページ数の実数調査が1998年頃には難しくなったことを示している．

そこで，検索エンジンを利用したWebページ数の推定方法を日本語のWebに用いることができるかどうか，用いることができた場合，どのような推定値を示すかを，実際に調査することにした．この推定方法が実際に適用できるかどうかは次の条件に大きく依存する．

- 対象とするWebから十分に多くのWebページを収集した検索エンジンが複数存在する．
- 適切なクエリによって十分に大きい標本集合を得ることができる．

次章では，Lawrenceらの推定方法の詳細，および信頼区間の計算方法について紹介する．この方法を日本語Webに適切に適用できたかどうか，つまり，我々の調査が英語のWebページ数の推定と同程度に信頼がおけるかどうかを信頼区間の大きさから判断した．

3. Webページ数の推定方法

3.1 対象とするWebページ

検索エンジンを利用したWebページ数の統計的推定では，全文検索システムのクエリに使用できるような文字列を最低1個は含むテキストページを対象とし，次のようなWebページは除外する．

- Firewallや Webサーバなどのアクセス制限の対象になっている．
- “robots.txt” を使って，クローラが収集しない設定になっている．

この集合のことを，Lawrenceらは publicly indexable Webと呼んでいる[1]．これ以降，このURL集合を U とし，$N \equiv |U|$ をWebページ数と呼ぶ．

3.2 用語の定義

まず，検索エンジンの集合 S とクエリの集合 Q を考える．検索エンジンの集合 S とは，goo，Lycosなどの，存在する検索エンジンすべての集合で有限である．クエリの集合 Q とは，これらの検索エンジンに与えることのできる検索文字列の集まりで，語を任意個数並べてよいと考えると，事実上 Q は無限集合である．

これらの集合の要素 $s \in S$，$q \in Q$ を使って，次の2種類のURLの集合を定義する：

$$U_s{}^q \equiv \{u|\ \text{クエリ } q \text{ を検索エンジン } s \text{ に与えると得られる検索結果中の URL } u\}$$

$$U_s \equiv \bigcup_{q \in Q} U_s{}^q.$$

つまり，U_s は検索エンジン s によって検索可能なURL

集合を表している．

3.3 確率の定義

全 URL 集合 U の任意の部分集合 X に対して確率 $P(X)$ を次のように定義する：

$$P(X) \equiv \frac{|X|}{|U|}.$$

このとき，事象 U_s である確率

$$P(U_s) = \frac{|U_s|}{|U|}$$

と検索エンジン s が持つ URL 集合の大きさ $|U_s|$ が分かれば，Web ページ数 $N = |U|$ は

$$N = \frac{|U_s|}{P(U_s)} \tag{1}$$

で求めることができる．

3.4 確率の推定方法

2 つの事象 A，B が，独立事象であれば，

$$P(A) = P(A|B) = \frac{|A \cap B|}{|B|}$$

が成り立つ．2 つの検索エンジン a と b のクローラが互いに独立して Web ページを収集していると仮定すると，U_a と U_b は独立した事象と見なせるので，次が成立する：

$$\begin{aligned} P(U_a) &= P(U_a|U_b) \\ &= \frac{|U_a \cap U_b|}{|U_b|}. \end{aligned}$$

このとき，$P(U_a)$ は $|U_a \cap U_b|$ / $|U_b|$ から求めることができる．しかし，検索エンジンは収集ページ数 $|U_a|$ や $|U_b|$ を公開しても，その内容 U_a や U_b を通常公開しない．そのため実際に $U_a \cap U_b$ を調べることは難しい．そこでクエリの有限部分集合 $Q' \subset Q$ を選び，

$$U'_a \equiv \bigcup_{q \in Q'} U_a^q$$

を定義する．U'_a は，有限個の クエリを用意すれば，実際に検索エンジンを使って観察することができる．また， Q' が Q からランダムに選ばれるならば，近似的に次の関係が成り立つ：

$$\frac{|U_a \cap U_b|}{|U_b|} \approx \frac{|U'_a \cap U'_b|}{|U'_b|}.$$

このとき，$P(U_a)$ は次のように推定できる：

$$P(U_a) \approx \frac{|U'_a \cap U'_b|}{|U'_b|}.$$

これを式 (1) に使って N を得る．また，$P(U_b)$ からも，同様に N を得ることができる．そこで，2 つの N の平均を，検索エンジン a と b から得られる N の推定値とする．

3.5 区間推定

N ページの中から非復元的にランダムに $n = |U'_b|$ ページ取り出したとき，n のうちの X ページが U_a に属する X の確率分布は超幾何分布となる[12) p.109–111]．$n \ll N$ の場合，1 ページを取り出す結果は次の 1 ページを取り出す結果にほとんど影響しない．したがって，$p = P(U_a)$ とおくと X の確率分布は 2 項分布 $B(n, p)$ となる．このとき，期待値 $E(X)$ と分散 $\sigma^2(X)$ は次のようになる．

$$E(X) \approx np$$
$$\sigma^2(X) \approx np(1-p)$$

n が大きいと，中心極限定理により X の確率分布は正規分布に近づく[12) p.170]． このとき，

$$\overline{p} = \frac{|U'_a \cap U'_b|}{|U'_b|}$$

を p の観測値とすると p の 95%信頼区間は近似的に

$$\left[\overline{p} - 1.96\sqrt{\overline{p}(1-\overline{p})/n},\quad \overline{p} + 1.96\sqrt{\overline{p}(1-\overline{p})/n}\right]$$

で求めることができる．

4. 実験方法

次の 3 検索エンジンを今回の調査に使用することにし，以下にのべる手順で実験を行った．

- goo `http://www.goo.ne.jp`
- Lycos 日本版 `http://www.lycos.co.jp`
- Infoseek 日本版 `http://www.infoseek.co.jp`

3 検索エンジンのほかに，AltaVista，Ring，Excite 日本版などを調査に利用することを検討したが，検索結果が安定しない，無効 URL 率が高い，収集 Web ページ数を公表していないなどの理由で利用しなかった．

手順 1：URL 集合の取得とクエリの選定

まず，1994，95 年の 2 年間の毎日新聞記事を形態素解析し，英数字を含まず，ひらがな，カタカナ，漢字のどれか一種からなる名詞，約 143,461 語を選んだ．英数字を含む語を除いた理由は，日本語を含むページを検索対象とするためである．かな漢字混じり語を除いた理由は，検索エンジンによる語の分割の可能性を低くするためである．

このようにして選んだ語を上記の 3 検索エンジンで検索し，その検索結果として URL 集合を得た．この URL 集合を調査し，次の 3 条件にあてはまる 291 語を調べ，これを有限クエリ集合（Q'）として採用した．

a. 各検索エンジンでの検索結果の URL 数が 50 個以上である．

表 1 日本語 Web ページ数の推定値
Table 1 Estimated number of Japanese Web pages.

a	b	$\lvert U'_a\rvert$	$\lvert U'_b\rvert$	$\lvert U'_a \cap U'_b\rvert$	$\overline{p}$	p の 95%信頼区間	ページ数の推定値 N
goo	Lycos	17,125	44,817	4,047	0.090	0.088 ～ 0.093	256,000,000
Lycos	goo	44,817	17,125	4,047	0.236	0.230 ～ 0.243	
goo	Infoseek	17,125	65,893	5,721	0.087	0.085 ～ 0.089	230,000,000
Infoseek	goo	65,893	17,125	5,721	0.334	0.327 ～ 0.341	
Lycos	Infoseek	44,817	65,893	21,890	0.332	0.329 ～ 0.336	59,300,000
Infoseek	Lycos	65,893	44,817	21,890	0.488	0.484 ～ 0.493	

b. 3 検索エンジンの検索結果の和集合の大きさが 600 個以下である．

条件 a の下限は，ある検索エンジンでインデックスの対象にまったくならないクエリを採用しないために設定している．50 という値は Lawrence らの調査と一致させるために採用した．条件 b は，単独のクエリの検索結果が大きな影響を与えない役割を果たしている．また，条件 b により，各検索エンジンでの検索結果件数も 600 個以下になるが，Infoseek については表示 URL 数の上限から 500 個以下にしている．

手順 2：URL の正規化

次に，上記の URL 集合の各要素である URL 名に以下の正規化を行い，重複する URL 名を除外した．

(1) ホスト名の小文字化
(2) ホスト名の後の `:80` の除去
(3) 16 進数表現文字の変換（例：`%7E` を `~` にする）
(4) index.html の除去して `/` で終わる URL として統一する．

手順 3：ページ集合の取得

上記で得られた URL すべてについて，GET 要求を出すことにより，その URL に対応する Web ページが実際に存在するかを調べた．調べた期間は 2000 年 10 月 1 日から 3 日の間で，ページの存在が確認できなかった URL を無効 URL として URL 名の集合から除いた．

手順 4：クエリの存在確認

GET 要求で取得できた Web ページの内容に，クエリに該当する文字列が含まれるかどうかを調べた．perl の文字列パターンマッチ機能を使い，空白文字，改行文字と中黒「・」を含むクエリがあればクエリを含むページとし，そうでない場合はクエリを含まないページとした．

手順 5：重複ページ集合の生成

最後に，3 検索エンジンから互いに異なる 2 検索エンジンを選び，それら 3 組について重複集合 $U'_a \cap U'_b$ を求めた．また，3 検索エンジンで検索できたページの和集合も求めた．

5. 実験結果

5.1 実験結果 1：日本語 Web ページ総数

実験の手順 5 で生成した重複集合を利用して，現在（2000 年 10 月 1～3 日）の日本語 Web ページ総数 N について，**表 1** に示す推定値を得た．この表では，3 検索エンジンが公表した収集ページ数をもとに N の推定を行っている．goo は 3,500 万，Lycos は 3,000 万，Infoseek は 1,800 万と収集ページ数を公表している[10),11)]．

表 1 の N の欄の推定値が一致しない原因はいくつか考えられる．まず，重複集合の占める割合 $\overline{p}$ のばらつきである．これは，クローラの収集する動作の独立性にばらつきがあることを示す．次に検索エンジンが公表したページ数に含まれるページの範囲の違いが考えられる．たとえば，クローラが収集したページすべてを数えるか，クエリに利用できるような文字列を最低 1 個含むページだけを数えるかでページ数が異なる．

今回の調査では，Lawrence らの方法にならって，公表したページ数の大きい組合せである goo と Lycos から推定した，256,000,000 個を日本語 Web ページ数として採用することにした．6 章でこの組合せを基に推定する理由についてさらに考察する．

5.2 実験結果 2：無効 URL と無効ページ

手順 3 で除いた URL は，次のような原因で対応する Web ページを取得できなかった URL で，これらを無効 URL と呼ぶ．

(a) connect できない，80 秒で time out したなどの理由で Web サーバの存在が確認できなかった．
(b) サーバからページが存在しないと応答があった．

また，手順 4 で除いたページは，存在したがクエリを含まなかったページで，これらを無効ページと呼ぶ．無効ページの原因は大まかには次の 2 つである．

(c) 新聞記事，電子掲示板，日記など，頻繁に書き換えられるページであるため，検索エンジンがインデックスを生成したときと内容が異なる．
(d) クエリを分割して複合語検索をしたり，関連語も検索対象にしたりしたため，単純なパターンマッ

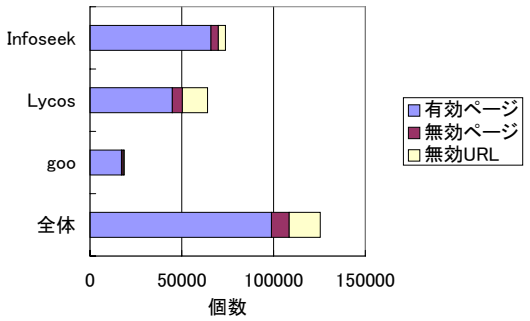

図 1 各検索サービスにおける無効ページと無効 URL 数
Fig. 1 Invalid pages and invaild URLs.

チで クエリの存在を確認できなかった．

無効 URL や原因 (c) の無効ページが多いのは，検索エンジンのクローラの収集頻度が低いためと考えられる．**図 1** は，検索結果の URL 集合の大きさと，その中に無効 URL と無効ページがどれくらいあったかを示す．なお，今回の調査では，無効ページの原因 (c) と (d) を区別していない．この図から，無効 URL と無効ページの存在する率は goo が最も低く，最も頻繁に収集を行っている可能性が高いことが分かる．また，他の 2 検索エンジンでも，無効 URL と無効ページを合わせて 2 割程度で，Web の更新状況を比較的よく反映している[13]．

図 1 は，使用したクエリ集合に対する検索結果件数は，Infoseek が最も大きいことも示している．これは 3 検索エンジンの中で Infoseek の収集ページ数が最も少ないことと矛盾しておらず，Infoseek は検索結果件数が 500 個以下の範囲におさまるクエリの存在確率が，他の 2 検索エンジンより高いことが原因だと考える．つまり，4 章で示した条件 b を適用せず，600 個以上の検索結果件数になるクエリも含めた調査ができれば，実際に収集したページ数やインデックスの大きさに近い大小関係になっただろうと考える．

6. 考　　察

6.1 推定値とモデルの妥当性

検索エンジンの検索結果の重複を利用した推定方法を日本語 Web ページに適用した結果，Lawrence らの仮定したモデルが正しいとすると，2000 年 10 月の時点で日本語の Web ページは最低 2 億 5,600 万個であると推定できた．95%信頼区間は約 ±600 万ページで，この大きさから，Lawrence らが行った調査程度には十分な大きさの標本集合を調査したことが分かる．

実際の Web に Lawrence らの仮定したモデルがあてはまらない要因を以下に考察する．

6.1.1 クローラの動作の独立性

検索エンジンのクローラがページを収集するとき，次のようなページを優先して集めることがある．

- ユーザが検索対象にしてほしいと検索エンジンに登録したページ．
- 人気のある Web サイト（例：Yahoo ディレクトリ）からリンクされているページ．
- 被リンク数の多いページ．特にページの属する Web サイト以外からの被リンク数の多いページ．
- URL 名のパス名部分が短い（Web サーバのルートドキュメントに近い）ページ．
- 更新頻度が高いページ．

このようなページを優先して集めた場合，異なるクローラが収集したページ集合の間の重複集合は，ランダムに集めた場合の重複集合より大きくなる．しかし，上記の条件にあてはまるページ数は限られているので，集めたページが多くなるほど，優先して集めたページが重複集合の中に占める割合が小さくなると考えられる．つまり，集めたページ数が大きくなるほど，片寄りのある場合からランダムに集めた場合に近づくと考えられる．実際，Lawrence らの調査でも，我々の調査でも，検索結果件数の中に占める重複集合の割合は，公表した収集ページ数が大きい 2 検索エンジンの間で最も低い．

そこで，我々の調査でも公表した収集ページ数の大きい goo と Lycos の組合せを利用して推定することにした．仮にこれらのクローラが互いに独立して収集していないとしても Web ページ数の推定値を小さくする方向に影響するので，クローラの独立性は Web ページ数の「下限」の推定には影響しないと考える．

6.1.2 インデックスの作成方式

収集したページ集合が同じでクエリが同一であっても，インデックスの作成方式が同じでないと検索結果が異なることがある．

4 章の条件 a により各検索エンジンでの検索結果の URL 数がゼロでないことが保証されるので，検索エンジンのどれかが完全に無視する文字列を除外している．

条件 b では和集合の大きさに制限を加えて，各エンジンのクエリの取扱いが大きく異なる場合をある程度除外している．それでも，使用したクエリには，表記のゆれの影響を受ける語がいくつかある．これは，4 章の手順 1 で複合語検索を起こさないためにつけた制約によって，クエリにカタカナ語が多くなったためと考えられる．たとえば「キャフェ」で検索した場合，「カフェ」を含むページの URL が検索結果に含まれる可能性がある．そこで，表記のゆれや関連語への対

応をしている検索エンジンの検索結果には，元のクエリを含まないページが多数含まれる可能性がある．検索結果の URL に対応するページの内容にクエリに使用した文字列が含まれることを調べ，含まれない場合は無効ページとして，Web ページ数の推定に利用していない．また，無効ページ数が検索結果に含まれる割合は，図 1 に示すように，最も高い検索エンジンでも 8.6%であった．そこで，今回の調査結果は関連語や表記のゆれの影響を受けていたとしても，約 10（≈ 8.6/91.4）%程度であろうと考える．

一方，表記のゆれや関連語検索とは逆に，あるクエリ q がページ p に含まれているにもかかわらず，クエリ q の検索結果に p が含まれていないことがある．たとえば，我々が 1999 年 10 月の調査で使用した 597 語を並べたページを Web 上で公開しておいたところ，ある検索エンジンのクローラで収集され，その検索エンジンで実際に検索できるようになった．しかし，597 語全部では検索できず，約 240 語のみで検索できた．このように検索できない クエリ q が存在する理由は，次のようなものが考えられる．

- クエリ q より重要な検索語がページ p に非常に多く含まれていたので，q がインデックスの対象から除外された．
- 検索エンジンがページの先頭から一定の範囲内にある語だけをインデックスの対象にしており，クエリ q がページの後ろの方にあるので，インデックスの対象から除外された．

このようなクエリ q が，今回の調査で使用したクエリの中に多く存在すると検索エンジンで収集されているにもかかわらず検索結果に反映されないページが多くなる．そのため，調査した重複集合が実際の重複集合より小さくなり，Web ページ数の推定値が実際より大きくなる．そこで，このようなページ，重要度の低いクエリ q を使用したために検索できなかったページは少ないことを次の方法で確認することにした．

goo と Lycos の検索結果から，片方のみから検索できたページで，サイズの大きいページを 10 個選ぶ．次に，そのページを眺め，そのページに含まれる語でクエリに使用するのが妥当だと思われる語を選び，それらを使って検索し，検索結果件数が 600 件以下の 5 語を選んだ．5 語の検索結果に，10 個のページは含まれていなかった．つまり，検索できなかった方の検索エンジンから新しく検索できるようになったページはなかった．

本来は，上記の方法ではなく，今回の調査で標本集合として採用した約 9 万 8,000 ページについて，各検索エンジンで収集されているかどうかを調査すべきだが，各検索エンジンはそのような手段を提供していないので，上記の方法で代用した．

6.1.3 公表しているページ数の範囲

この Web ページ数の推定では，検索エンジンが公表している収集ページ数を利用しているので，収集したページの意味によって日本語 Web ページ数の大きさが異なる．通常，クローラが収集したページには，インデックスの対象にならない次のようなページが含まれていることが多い．

- ファイルとしては存在するが，タグのみを含み，テキストを含まないページ
- テキストを含むが “Under Construction” のような非常によく使われるテキストであるためインデックスの対象とならない語のみを含むページ
- 日本語と英語以外のテキストのみを含むページ
- すでに収集したページと内容が同一と見なせる（ミラー）ページ

各検索エンジンが上記のページを含んだ数をページ数として公表しているのであれば，我々の推定した数は，上記のページを含んだ数になる．実際には，各検索エンジンは公表ページ数を「登録 URL 件数」「日本語サイト 3,500 万 URL」などと記しているので，上記のページはある程度除き，データベースに登録できたページ数を公表していると考える．一方，通信白書の採用した方式ではクローラで集めたページをすべて数えており，今回の調査対象であるページ集合より広い範囲のページを含むページ集合を調査した推定値であると考える．

6.1.4 各要因の影響

上記の 3 要因が大きな影響を持つと，本研究で調査した推定値は Web ページ数の下限値として小さすぎる可能性と大きすぎる可能性の両方がある．

次の要因の影響が大きいと小さすぎる推定値になる．

- クローラの動作に独立性がない．
- 検索エンジンが実際より少ない収集ページ数を公表している．

この研究で推定しようとしているのは日本語 Web ページ数の「下限」なので，小さすぎる推定値になるのはやむをえないと考える．

一方，次の要因が大きな影響を与えていると大きすぎる推定値になる．

- インデックスの対象にならないクエリを多く使用した．
- 検索エンジンが実際より大きい収集ページ数を公表している．

最初の要因によって大きすぎる推定値である場合は1，2 割程度の過大評価であり，その分を考慮しても，通信白書の推定値よりかなり大きいといえる．**図 2** は，今回および過去に我々が同じ推定方法を使用して得た推定値[14),15)]と，通信白書の推定値の関係を示す．この図から，日本語の Web ページ数は通信白書の示す値以上に急速に増大していることが分かる．

6.2 経時変化の測定

Lawrence らの方法を日本語 Web に今回適用できたのは，彼らが使用した条件と同じ条件を満たす日本語のクエリを十分な数見つけることができ，その結果十分な大きさの標本集合を作ることができたためである．しかし，この方法を使った調査を今後も継続して経時変化を観察するには次のような問題が考えられる．

まず，第 1 の問題点として，調査に利用できる日本語検索エンジンの少なさがあげられる．この調査に利用する日本語検索エンジンは次の条件を満たす必要がある．

- 収集ページ数が多く，ページ数を公表している．
- 更新頻度が高い．
- サービスを安定して提供している．

この条件にあてはまる検索エンジンは今回の調査では Lycos，goo，Infoseek の 3 サービスしかなく，そのうち，検索結果件数に制限がないのは 2 サービスであった．今後，上記の条件に適合する日本語検索エンジンが多くできることが期待される．

第 2 の問題点としてクエリの選定の難しさがあげられる．我々はこの手法を日本語の Web に過去数回適用しており[14),15)]，**表 2** は，その適用時期，使用したクエリ集合，推定した N の値を示す．この表は，日本語の Web ページ数の増加とともにクエリ集合が小さくなり，クエリ集合の出典を変えたことを示している．これは，ある調査でクエリが検索結果件数の和集合の大きさ 600 個以下の条件を満たしても，次の調査では 600 個を超えることが多く，クエリに要求される条件を満たさなくなためである．つまり，現在の方法

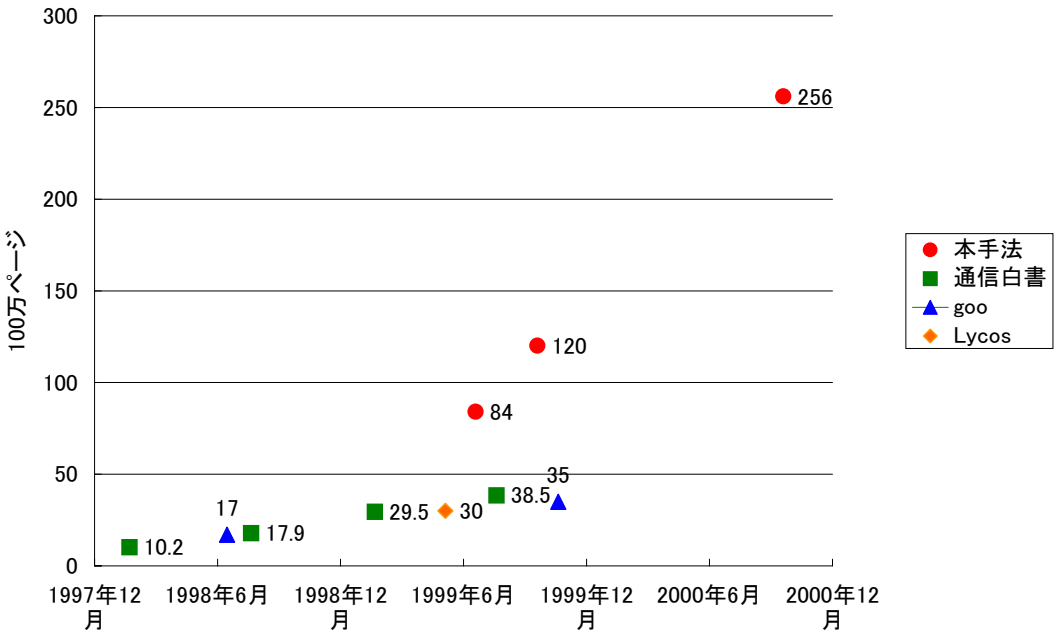

図 2 日本語 Web ページ数の推定数の変化
Fig. 2 Growth of estimated number of Japanese Web pages.

表 2 日本語 Web ページ数の推定に使用したクエリ数の変化
Table 2 Queries used in the search engine coverage overlap method.

推定時期	ページ数推定値 N	使用したクエリ数	クエリの出典
1999 年 7 月	84×10^6	1,085	検索エンジンの利用履歴
1999 年 10 月	120×10^6	597	現代用語の基礎知識'99 の見出し語
2000 年 10 月	256×10^6	291	毎日新聞記事 (1994，1995)

は調査のたびにクエリを毎回選定しなおす必要があり，将来も適切なクエリを見つけ出せるかどうか分からない．この問題点を解決するには，600 個以下の条件をゆるめて 600 個より多くの検索結果件数になるクエリを使用することが考えられる．一方，第 1 の問題点で指摘したように，600 個より多くの検索結果件数を得ることのできる日本語検索エンジンは現状は 2 サービスしかない．調査に利用できるクエリ数の拡大のためにも日本語検索エンジンの充実が望まれる．

第 3 の問題点として，検索エンジンの設計方針の変化が考えられる．現在，英語圏の検索エンジンは大量の検索結果を表示するのではなく，少数でも有用な検索結果を表示する方向への変化が始まりつつある．たとえば “Internet” のような非常に多くのページに含まれるクエリについては同一サイトからは一定数のページしか表示しない，かなり広い範囲のページをミラーと見なして除外するという工夫が行われつつある．今回の調査に使用した日本語の検索エンジンではそのような傾向を観察できなかったが，この傾向が強くなると，検索エンジンにクエリを与えて Web ページの標本集合を得る考え方自体が推定に向かなくなる．

7. まとめと今後の方向

検索エンジンを利用した Web ページ数の統計的推定を日本語 Web に適用した結果，日本語 Web には，2000 年 10 月に少なくとも約 2 億 5,600 万ページ存在することが推定できた．これは，日本で最大の検索エンジンでも日本語 Web の約 13%しか検索できないことを意味する．一方，検索エンジンを使った推定方法は，長期間の経時変化の測定や，世界全体の Web ページ数の推定にあまり適していないことも分かった．

なお，検索エンジンを利用した推定方法の限界を解決するために，Lawrence ら自身も検索エンジンを利用しない Web ページ数の推定方法を後日発表し，1999 年 2 月の世界の Web ページ数は 8 億個であるとした[5]．この方法は IP アドレスをランダムに抽出して行うものだが，問題点がいくつかある．たとえば，JP ドメインに同じ方法を適用してみると約 1,900 万ページになる[16]．また，複数ホスト名の同一 IP アドレスへの割当てや IPv6 の導入などの影響により IP アドレス分布は今後大きく変化すると予想されるので，長期的な観測には Lawrence らの新しい方式もあまり適していない．Web に関する統計的推定には，さらに新しい方式の開発が今後必要と考える．

謝辞 本研究の一部は文部省による東京情報大学ハイテクリサーチセンタ助成と津田塾大学ハイテクリサーチセンタ助成によって行われた．

参 考 文 献

1) Lawrence, S. and Giles, C.L.: Searching the World Wide Web, *SCIENCE*, Vol.280, pp.98–100 (1998). http://www.neci.nj.nec.com/homepages/lawrence/websize.html
2) Berners-Lee, T.: *Weaving the Web*: *The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*, HarperCollins (1999).
3) Bray, T.: Measuring the Web, *Fifth International World Wide Web Conference* (1996). http://www5conf.inria.fr/fich_html/slides/papers/PS3/P9/T01.htm
4) Bharat, K. and Broder, A.: A technique for measuring the relative size and overlap of public Web search engines, *7th International World Wide Web Conference* (1998). http://www7.scu.edu.au/programme/fullpapers/1937/com1937.htm
5) Lawrence, S. and Giles, C.L.: Accessibility of information on the web, *NATURE*, Vol.400, pp.107–109 (1999). http://www.wwwmetrics.com
6) 総務省郵政事業庁：通信白書平成 12 年版 (2000). http://www.mpt.go.jp/policyreports/japanese/papers/h12/html/data-1-index.html
7) 外薗博文：日本のインターネット（WWW）の現状，郵政研究所月報 **9**, pp.79–86 (1998).
8) 宮沢 浩：日本のインターネット（WWW）の現状その 2，郵政研究所月報 **12**, pp.99–102 (1998).
9) エヌ・ティ・ティ・アド：ニュースリリース 1998 年 6 月 26 日．http://www.goo.ne.jp/help/info/n_release/n_980626.html
10) エヌ・ティ・ティエックス：ニュースリリース 2000 年 10 月 3 日．http://www.goo.ne.jp/help/info/n_release/n_001005.html
11) Lycos Japan: Lycos Japan がサイト全般に渡る大規模リニューアルを実施．http://www.lycos.co.jp/help/info/release.html?press=06
12) 東京大学教養学部統計学教室（編）：統計学入門—基礎統計学 I，東京大学出版会 (1991).
13) Brewington, B.E. and Cybenko, G.: How dynamic is the Web? *Proc. 9th International World Wide Web Conference*, pp.257–275 (2000).
14) 大森貴博，笹塚清二，近藤晶子，水谷正大，来住伸子，小川貴英：統計的手法による日本語 Web の調査，情報処理学会第 59 回全国大会論文集，3P-01 (1999).
15) 来住伸子，大森貴博，笹塚清二，近藤晶子，水谷正大，小川貴英：統計的推定による日本語 Web

の調査，インターネットコンファレンス'99 論文集，pp.21–28 (1999).
16) Kishi, N., Ohmori, T., Sasazuka, S., Kondo, A., Mizutani, M. and Ogawa, T.: Estimating Web Properties by Using Search Engines and Random Crawlers, *Proc. INET 2000*, Internet Society (2000). http://www.isoc.org/inet2000/cdproceedings/2a/2a_3.htm

(平成 12 年 12 月 20 日受付)
(平成 13 年 4 月 13 日採録)

（担当編集委員　加藤 和彦）

来住　伸子（正会員）
1983 年東京大学大学院工学系研究科情報工学専攻修士課程修了．同年日本アイ・ビー・エム（株）入社．1992 年より津田塾大学助教授．

大森　貴博（学生会員）
2001 年東京情報大学大学院博士課程中退．2001 年より（株）オン・ザ・エッヂ勤務．

水谷　正大（正会員）
1986 年早稲田大学大学院理工学研究科物理および応用物理学専攻博士課程修了．理学博士．2001 年より東京情報大学教授．

小川　貴英（正会員）
1974 年東京大学大学院工学系研究科情報工学専攻修士課程修了．東京大学工学部助手を経て，現在津田塾大学教授．