統計的推定による日本語 Web の調査

Statistical Analysis of Web Pages in Japanese and in Japan

来住 伸子 [*1] 大森 貴博 [*2] 笹塚 清二 [*2] 近藤 晶子 [*1] 水谷 正大 [*2] 小川 貴英 [*1]
[*1] 津田塾大学 数理情報科学科, [*2] 東京情報大学 情報学科

Nobuko Kishi[†1] Takahiro Ohmori[†2] Seiji Sasazuka[†2] Akiko Kondo[†1]
Masahiro Mizutani[†2] Takahide Ogawa[†1]
[†1]Department of Mathematics and Computer Science, Tsuda College,
[†2]Department of Information Systems , Tokyo University of Information Sciences
{kishi,m99kondo,ogawa}@tsuda.ac.jp, {ohmori,sasazuka,mizutani}@rsch.tuis.ac.jp

概要 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 らの提案した推定方法によって、日本語の Web ページ数の統計的推定を行なった。その結果、1999 年 11 月現在、少なくとも 約 1 億 2000 万の日本語の Web ページ存在することが推定できた。さらに、日本国内の 4 検索エンジンについて、インデックスの相対的な大きさや無効 URL の存在率などの比較を行なった。また、今回の方法以外の統計調査方法、たとえば、クローラーを利用した方法についての考察を行なった。

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、検索エンジンで、どのくらいの数の Web ページを実際に検索できるのか、について、正確なデータは現状ではあまり無い。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。

最近、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]。そこで、この推定方法を 日本語の検索エンジンに対して使うことにより、日本語の Web ページ数の推定を行なった。また、この推定方法の統計的信頼性についての詳細な検討も行なった。この推定により、1999 年 11 月の日本語の Web ページは最低 約 1 億 2000 万ページあ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、日本語の Web ページ全体の約 29 % しか検索できないことを示す。また、
Lawrence らの検索エンジンを利用した推定方法は、統計的信頼性は非常に高いことも分かった。

本稿では、第 2 節 で、 従来の Web の調査方法をいくつか紹介する。特に、この研究で使用した Lawrence らの検索エンジンを利用した調査方法を紹介する。第 3 節 では、Web ページ数の推定を中心に、統計的推定の詳細な説明を行なう。つづいて、第 4 節では、実際の実験方法を述べ、第 5 節で, その実験結果を紹介する。実験結果として、日本語 Web ページの数の他に、検索サービスの URL 数 の相対比較と無効 URL 存在率も紹介する。第 6 節 では実験結果を考察し、さらに、最近の動向として検索エンジンを利用しない推定方法を考察する。

2 従来の研究

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 サーバー数が増加した [2, 3]。1993 年から 2 年間程は、 Web ページ数がまだ大きくなかったため、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、数種の統計量について報告を行なっている [4]。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もつ URL 集合の中で、重複する URL 集合を利用して、Web ページの調査を行なうことを提案した [5]。しかし、Bharat らの方法は、集合和の計算方法 等に問題があり、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 クエリーを選び, それらを AltaVista, Excite, HotBot, Infoseek, Lycos, Northern Light の 6 検索サービスに与えた。検索結果として返ってきたそれぞれの URL 群から、各検索サービス間の URL の重複を計算し、検索可能な公開された Web ページ数は, 1997 年 12 月には 最低 3 億 2000 万ページであったと推定した。

日本では、平成 11 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 万ページという数値が記載されている [8]。これらの数値は、郵政省郵政研究所による調査 [6, 7] 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 [14, 15]、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 [16] を収集したことを公表している。しかし、これらの検索エンジン単独で 日本語 Web ページ全てを検索できないのが現状である。そこで、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。

3 Web ページ数の推定方法

3.1 対象とする Web ページ

Web ページ についての統計的推定では、全文検索システムでの クエリーに使用できるような文字列を最低 1 個は含む テキストデータを対象とし、次のような Web ページは除外する。

- Firewall や Web Server などのアクセス制限の対象になっている。
- “robot.txt” を使って、web ロボットが収集しない設定になっている。

この集合のことを、[1] は publicly indexable web と呼んでいる。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

いま、次の集合を考えよう：

$$\begin{aligned} S &: \text{検索エンジンの集合} \\ Q &: \text{クエリーの集合} \end{aligned}$$

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。クエリーの集合 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合で、語を任意個数並べてよいと考えると、大きさ無限の集合である。

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。

$$\begin{aligned} {U_s}^q &\equiv \{u| \text{クエリー } q \text{ を検索エンジン } s \text{ に与えると検索結果として得られる URL } u\} \\ U_s &\equiv \bigcup_{q \in Q} {U_s}^q \end{aligned}$$

つまり、U_s は 検索エンジン s によって検索可能な URL の集合を表している。

3.3 確率の定義

全 URL 集合 U の任意の部分集合 $X \subseteq U$ 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、

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

により、 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 $N \equiv |U|$ 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

二つの事象 $A, B \subseteq U$ が、独立事象であれば、

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

が成り立つ。もし、二つの検索エンジン a と b の Web クローラーが互いに独立して Web ページ を収集したとすると、U_a と U_b は, 独立した事象と考えられる。このとき、

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

が成立する。よって、$\frac{|U_a \cap U_b|}{|U_b|}$ により、$P(U_a)$ を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、クエリーの有限部分集合 $Q' \subset Q$ 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 クエリーを用意すれば、実際に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、U'_a が U_a からランダムに選ばれるならば、次の関係がなりたつ。

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

そこで、$P(U_a)$ を次のように推定する。

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

3.5 区間推定

N ページの中から非復元（without replacement）でランダムに $n = |U'_b|$ ページ取り出したとき、n のうちの x ページが U_a に属している確率 $f(x)$ は、超幾何分布となる [17, 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、$p = P(U_a) = \frac{M}{N}$ とおくと、$f(x)$ は $Bi(n, p)$ の 2 項分布と同様に振る舞う。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、$f(x)$ は、正規分布 に近付づく [17, p.170]。 $\overline{p} = \frac{|U'_a \cap U'_b|}{|U'_b|}$ を p の観測値 とすると、p の 95 % 信頼区間は、近似的に

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

で求めることができる。

4 実験方法

次の 4 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、実験を行なった。

- goo[9]
- Excite 日本語版 PowerSearch[10]
- Lycos 日本語版 [11]
- Infoseek 日本語版 [12]

手順 1: URL 集合の取得とクエリーの選定

まず、『現代用語の基礎知識 1999 年度版』[13] から得た 17,737 語から、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 4 検索サービスで検索し、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 597 語を調べ、これを有限クエリ集合 (Q') として採用した。

- Infoseek の検索結果の URL 数が 50 個から 500 個の範囲, 他の 3 検索エンジンでの検索

結果の URL 数が 50 個から 600 個の範囲に入る。

- 各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。

和集合で 600 個 の上限は、実際に検索エンジンから、検索結果中の URL を実際に調べられる上限として設定した。また、単独の クエリーの検索結果が大きな影響を与えない役割も果たしている。50 の下限は、同一 クエリーが全く異なる検索結果になる場合を除くために設定した。

手順 2:URL の正規化

次に、上記の URL 集合の各要素、URL 名 に対して、次のような正規化を行なった。

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 全てについて、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 1999 年 11 月 11 日から 13 日の間で、ページの存在が確認できなかった URL と確認時に Time out した URL(今回の実験では 80 秒) を無効 URL として URL 群から除いた。

手順 4:クエリーの存在確認

GET 要求で取得できた Web ページの内容に、クエリー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 クエリーがあれば、クエリーを含むページとし、そうでない場合は、クエリーを含まないページとした。例えば クエリー「年金払積立傷害保険」に対して「年金払・積立傷害保険」を含むページは、クエリーを含むページとするが、「積立傷害保険を年金払にする」を含むページは、クエリーを含まないページとする。

手順 5:重複ページ集合の生成

最後に、4 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 6 組について、重複 ページ 集合、$U'_a \cap U'_b$ を作成した。また、4 検索エンジンで検索できたページ集合の和集合も作成した。

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 (1999 年 11 月 11–13 日) の日本語 Web ページ総数 N について、表 1 および表 2 のような推定値が得られた。表 1 では, 収集 ページ数を 3500 万と 公表 している検索サービス goo [15] を a , $|U_a| = 3.5 \times 10^6$ とし、他の 3 サービスを b として、N の推定を行なった。表 2 では、収集 ページ数を 3000 万と 公表 している検索サービス lycos [16] を a , $|U_a| = 3.0 \times 10^6$ とし, 他の 3 サービスを b として、N の推定を行なった。

表 1 と表 2 の N の欄で示す 6 個の推定値が一致しない理由は、いくつかある。まず、表 1 と表 2 の違いは、goo と lycos の URL 数の数え方の違いのためと考えられる。[18] に記述してあるように、クローラーが収集したページ全てを数えるか、クエリー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 URL 集合の大きさが異なる。

また、クローラー の収集するページ集合の独立性が低いと、推定値に誤差を生じる。最近のクローラーは、人気が高いと推定できる URL から優先的に集める技術を使用することがある。そのため、小さな検索エンジンでは人気の高いページが占める割合が高くなり、他の検索エンジンとの重複が大きくなる。そこで、他の検索エンジンとの重複が小さい、つまり p が小さい検索エンジンの組み合わせの方が、互いの独立性が高いと考えられる。したがって、2 個の表から、p が最も低い組み合わせ、goo と lycos から推定した値,120×10^6 を日本

表 1: goo を元にした日本語 Web ページ数の推定

b	$\lvert U'_b \rvert$	$\lvert U'_a \cap U'_b \rvert$	$\overline{p}$	N	95% 信頼区間
Excite	55,395	17,232	0.311	113×10^6	$111 \sim 113 \times 10^6$
Infoseek	81,669	21,079	0.258	136×10^6	$134 \sim 137 \times 10^6$
Lycos	60,158	15,533	0.258	136×10^6	$134 \sim 137 \times 10^6$

表 2: Lycos を元にした日本語 Web ページ数の推定

b	$\lvert U'_b \rvert$	$\lvert U'_a \cap U'_b \rvert$	$\overline{p}$	N	95% 信頼区間
Excite	55,395	22,472	0.406	74.0×10^6	$73.2 \sim 74.7 \times 10^6$
goo	61,888	15,533	0.251	120×10^6	$118 \sim 121 \times 10^6$
Infoseek	81,669	27711	0.340	88.4×10^6	$87.6 \sim 89.3 \times 10^6$

語 Web ページ数として採用し、日本語 Web ページ数は、少なくとも約 1 億 2000 万であると推定する。

5.2 実験結果 2 : インデックス の相対的な大きさ

4 検索エンジンで検索できたページ集合の和集合の大きさ 162,041 を 1 とした時の、各検索エンジンで検索できた Web ページ数 を表 3 と図 1 に示す。これにより、今回の調査対象とした、4 検索エンジンでは、Infoseek が最大 の検索用インデックスを持つ結果となる。一方、Infoseek Japan に問い合わせたところ、保有ページ数は 1800 万なので、goo や lycos より保有ページ数が少ない。つまり、検索エンジンの 保有ページ数とそのインデックスの大きさには、必ずしも比例関係が成り立たないことを示している。

比例関係が成り立たない原因は、同一のページに対し、検索エンジンが異なると、 異なる大きさのインデックスを作成するためであると考える。たとえば、あるページに「A, B, C, D」の 4 語が含まれていても、「A, B, C」の 3 語だけをインデックスに入れ、残り「 D」をインデックスに入れないという動作をする可能性がある。具体的には、今回の調査に使用した クエリー 597 語すべてを含む Web ページは、既にある検索エンジンのクローラーによって収集されており、その検索エンジンで実際に検索できる。しかし、クエリー 597 語 のどの語でも検索できるのではなく、一部の語でしか検索できない。Infoseek Japan の場合、保有ページ数が少ないが、各ページからインデックスに入れる語数が、他の検索エンジンより多いので、インデックスが最も大きくなったと考える。

表 3: インデックスの相対的な大きさ

	検索ページ数	相対的な大きさ	95% 信頼区間
Infoseek	83357	0.502	0.496∼0.509
goo	65736	0.381	0.375∼0.387
Lycos	62716	0.370	0.364∼0.376
Excite	57359	0.341	0.335∼0.347

5.3 実験結果 3 : jp ドメイン外のページの占める割合

今回の調査では、各検索エンジンには、検索対象のドメイン名に制限をつけずに検索を行なった。そのため、クエリーを含むページであれば、jp ドメイン外のページも検索結果に含まれる。実際に、jp ドメイン外のページが検索結果にどの程度の割合で含まれたかを表 4 に示す。この表から、 jp ドメイン外のページの占めるページの割合が検索エ

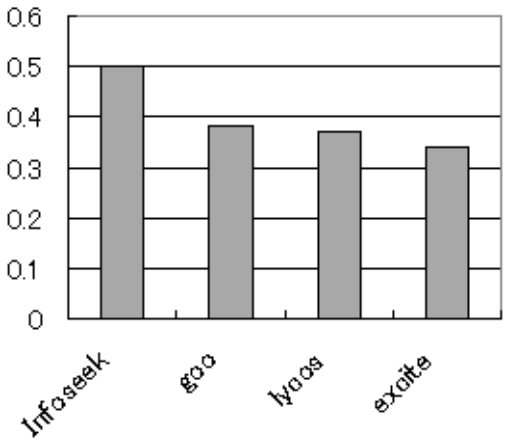

図 1: インデックスの相対的な大きさ

ンジンによって大きく異なることが分かる。

表 4: jp ドメイン以外のページが占める割合

	検索ページ数	.jp 外のページ数	.jp 外の割合 (%)
Infoseek	83357	7716	9.45
goo	65736	1222	1.97
Lycos	62716	2402	3.99
Excite	57359	4445	8.02
全体	162534	12459	7.67

5.4 実験結果 4: 無効 URL と無効ページ

手順 5 の 重複 URL 集合を計算する前の段階、手順 3 で除いた URL は、次のようなものであり、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 サーバーの存在が確認できなかった。

(b) ページの存在が確認できなかった。

また、手順 4 で除いたページは、ページは存在したが、クエリー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、大まかには次の二つである。

(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時と内容が異なる。

(d) クエリーを分割して AND 検索などをしたり、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チで クエリー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 Web クローラーの 収集頻度が低いためと考えられる。そこで、無効 URL が元の URL 群に占める割合を無効 URL 存在率、無効ページが元の URL 群に占める割合を無効 ページ 存在率とし、表 5 に示す。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。

この表では、Infoseek と goo の 2 検索エンジンが 無効 URL 存在率と無効ページ存在率の両方が低い。そこで、4 検索エンジンの中では、Infoseek と goo のクローラーの収集頻度が他の 2 検索エンジンより高い可能性が高い。

表 5: 各検索サービスにおける無効 URL 存在率

	無効 URL 存在率 (%)	無効ページ存在率 (%)
Infoseek	4.4	3.4
goo	4.9	2.4
Lycos	12.5	18.1
Excite	12.4	11.1
全体	10.4	11.8

6 考察と最近の動向

Lawrence らの方法を、日本語 Web ページに適用した結果、統計推定量としては高い精度で、日本語 Web ページ数を推定できることが確認できた。ただし、実際の 日本語 の Web ページ 数と一致しない要因としては、次のようなものが考えられる。

1. 検索エンジンの クローラーの動作の独立性

2. 検索エンジンのインデックスの作成方法
3. 検索エンジンが発表している URL 数の定義

これらの要因のうち、前者の二つに問題がある場合、つまり、独立性がない場合やインデックスの作成方法が不十分な場合は、この推定方法は、実数より小さい値を示すはずである。そこで、この研究の推定値、1 億 2000 万ページは、1999 年 11 月での 日本語 Web ページ数の最低推定値を与えると考える。これは、収集ページ数 を公表している検索エンジンの中では最大の値 3500 万を公表している goo でも、日本語の Web ページの約 29 % しか検索していないことを示す。一方、4 検索エンジンの検索結果の集合和の大きさと、goo の検索結果の集合の大きさの比から、4 検索エンジンすべてを利用すれば、日本語の Web ページ全体の約約 70 % が検索できることになる。

このように、この 研究で使用した推定方法は、統計推定における信頼区間という点では精度が高いが、検索エンジンのさまざまな特性の影響をうけるため、実際の 日本語 の Web ページ 数と必ずしも一致しない。また、クエリーの有限集合を利用しているので、その クエリーの性質の影響を受ける。日本語の クエリーのみを使うと日本語 の Web ページ数を推定し、英語の クエリーのみを使うと英語 の Web ページ数を推定することになる。つまり、我々の調査では、1999 年 11 月には 最低 1 億 2000 万の日本語の Web ページ、Lawrence らの調査では、1997 年 12 月には 最低 3 億 2000 万の英語 の Web ページがあったと推定している。

そこで、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 らは、検索エンジンを利用しない、Web ページ数の推定方法を最近発表し、1999 年 2 月の世界の、使用言語を限定しない Web ページ数は 8 億個であるとした [19]。

新しい推定方法は、次のような方法である。

1. IP アドレス空間 ($256^4 = 4.3 \times 10^9$) から、3.6×10^6 個の IP アドレスをランダムに選ぶ。
2. それらをサンプル調査することによって、269 個のアドレスに 1 個、Web サーバーを見つけた。これにより、
$$\frac{4.3 \times 10^9}{269} = 16 \times 10^6$$
個のサーバがあることを推定した。
3. しかし、これらは、名前の予約だけのページを含むので、それらを除く処理を行なった。その結果、世界の Web サーバーは、2.8×10^6 あると推定した。
4. ランダムに見つけた、Web サーバーの最初の 2500 個について、何個のページがあるかを調べた。その結果、1 サーバーあたり、平均 289 ページあることがわかった。
5. そこで世界の Web ページの数は、$2.8 \times 10^6 \times 289 = 8 \times 10^8$ とする。

しかし、この方法には、問題点がいくつかある。たとえば、日本に同じ方法を適用し、NetCraft 社 [20] の調査による jp ドメインと世界の Web サーバーの比率を利用して試算すると、jp ドメインには、以下の式から 約 1000 万ページしかないことになる。

$$8 \times 10^8 \frac{7}{540} = 10^7$$

また、IPv6 の導入により、国別や言語別の IP アドレス分布は大きく変化すると予想されるので、長期的な観測には、Lawrence らの新しい方式はあまり向いていない。

7 まとめと今後の方向

Lawrence らの方法 [1] を、日本語 Web に適用した結果、統計的推定としては高い精度で、日本語 Web には少なくとも 約 1 億 2000 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、日本語の Web ページの約 29 % しか検索していないことを示す。

今後、このような研究を継続して行なうことにより、日本語 Web ページ数や 日本語検索エンジンのインデックスの相対的な大きさなどの計時変化を観察することができる。英語 の Web では、Web の大きさの計時変化の調査はすでに行なわれているが [21, 22]、日本ではまだ行なわれていない。計時変化の調査から、日本語 Web の成長予想や日本語検索エンジンの性能予想が可能になり、新しい Web 応用技術開発や設備投資に有用な基礎データとなる。

また、世界の Web ページ数という、世界にとって重要なデータは、正確にはまだ推定されていない。世界の Web ページ数の推定は、Lawrence ら

の 最近の提案のように検索エンジンを使わない推定方法の可能性もあるが、対象言語別に、検索エンジンを使った推定を行なって集計するという方法も可能性がある。その他の方法も含めて、今後も継続した研究 [23] が必要である。

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なわれた。

参考文献

[1] Steve Lawrence, C.Lee Giles, *Searching the World Wide Web*, SCIENCE **280**, 99 (1998) http://www.sciencemag.org/cgi/content/abstract/280/5360/98, http://www.neci.nj.nec.com/homepages/lawrence/websize.html

[2] Robert Cailliau: *A Little History of the World Wide Web* http://www.w3.org/History.html

[3] Mathew Gray: *Measuring the Growth of the Web — June 1993 to June 1995* http://www.mit.edu/people/mkgray/growth/

[4] Tim Bray: *Measuring the Web,* Fifth International World Wide Web Conference May 1996 http://www5conf.inria.fr/fichhtml/slides/papers/PS3/P9/T01.htm

[5] Krishna Bharat and Andrei Broder: *A technique for measuring the relative size and overlap of public Web search engines*, Seventh International World Wide Web Conference, April 1998 http://www7.scu.edu.au/programme/fullpapers/1937/com1937.htm

[6] 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**, 79(1998)

[7] 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**, 99(1998)

[8] 郵政省 『平成 11 年版　通信白書』 http://www.mpt.go.jp/policyreports/japanese/papers/99wp/99wp-0-index.html

[9] http://www.goo.ne.jp

[10] http://www.excite.co.jp

[11] http://www.lycos.co.jp

[12] http://www.infoseek.co.jp

[13] 清水均編集「現代用語の基礎知識 1999」自由国民社、1999 年

[14] エヌ・ティ・ティ・アド『ニュースリリース: 350 万ページビュー／１日突破』 http://www.goo.ne.jp/help/n980626.html

[15] エヌ・ティ・ティエムイー情報流通『ニュースリリース: ポータルサイト “goo” の 検索機能を大幅に強化』 http://www.goo.ne.jp/help/n991005.html

[16] Lycos Japan,『プレスリリース：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/press06.html

[17] 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

[18] How to count URLs http://www.excite.com/ice/counting.html

[19] Steve Lawrence, C.Lee Giles, *Accessibility of information on the web*, NATURE **400**, 107 (1999) http://www.wwwmetrics.com

[20] Netcraft, *The Netcraft Web Server Survey,* http://www.netcraft.com/survey/

[21] Danny Sullivan, *Search Engine Sizes* http://searchenginewatch.com/reports/size.html

[22] Melee Productions, *Meleee's Index Engine Coverage Analysis* http://www.melee.com/mica/index.html

[23] 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 (平成 11 年後期) 講演論文集、3P-01, 1999.