3P－1

統計的手法による日本語Webの調査

大森 貴博* 笹塚 清二* 近藤 晶子† 水谷 正大* 来住 伸子† 小川 貴英†

* 東京情報大学 情報学科 ohmori@rsch.tuis.ac.jp, sasazuka@rsch.tuis.ac.jp, mizutani@rsch.tuis.ac.jp

† 津田塾大学 数理情報科学科 m99kondo@tsuda.ac.jp, kishi@tsuda.ac.jp, ogawa@tsuda.ac.jp

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) ページ数も急増している。実際にどのくらいの「Web ページ」が存在するか、そして、どの程度の増加率があるのかに関しての正確な統計データは現状ではあまり無い。本研究では、Web ページ数に関する統計が、Web に関する基礎研究として重要な意味を持つと考え、商用検索サービスを利用した日本語 Web ページ数の推定を行った。

2 従来の研究

全世界の Web ページ数の推定に関しては、各検索サービスが保有しているデータを利用した Lawrence らの報告 [1] がある。彼らの実験方法と、その結果は次のようである。

米国NECの従業員がWeb検索に使用したクエリから575個を抽出し、これらを AltaVista, Excite, HotBot, Infoseek, Lycos, Northern Light の 6 つの各検索サービスに与えて実験を行った。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、その比から Web 全体のページ数を推定した。URL が異なっても同一内容のページの場合は、重複として取り除かれた。その結果、検索サービスによって検索可能な公開されている Web ページ数は最低3億2000万ページと推定された。最近の発表 [2] では、8億ページと推定している。

日本では、郵政省郵政研究所による調査 [3, 4] があり、日本の Web ページ総数は「1000 万ページ (1998/ 2/10 ～ 2/26)」「1800 万ページ (1998/ 8/3 ～ 9/7)」と推定されている。

Measuring Japanese World Wide Web.

Takahiro Ohmori*, Seiji Sasazuka*, Akiko Kondo†, Masahiro Mizutani*, Nobuko Kishi†, Takahide Ogawa†

* Tokyo University of Information Sciences

† Tsuda College

3 推定方法

本研究では、以下のように対の検索サービス間の重複を利用した Lawrence らの推定方法にならって、日本語で書かれた Web ページ総数を推定した。

2 つの検索サービス a、b は互いに独立して一様に Web 上からデータを収集していると仮定する。あるクエリに関し、サービス b によって返される URL の数を N_b、サービス a と b の両方が重複して返す URL 数を N_{ab}、求める Web ページの総数を N としたとき、値 N_{ab}/N_b は N のうちサービス a が保有している Web ページ数の割合 P_a と等しいと考えられる：

$$P_a = N_{ab}/N_b$$

このとき、サービス a が保有している全 Web ページ数を S_a としたとき、

$$\frac{N_{ab}}{N_b} \approx \frac{S_a}{N}$$

が成立していると仮定すると、Web ページの総数 N は

$$N = S_a/P_a = S_a \cdot N_b/N_{ab}$$

と推定することができる。

4 実験方法

本研究の実験期間は 1999 年 7 月 2 日～7 月 13 日とし、次の検索サービスを使用した：goo (http://www.goo.ne.jp), Excite 日本語版 PowerSearch (http://www.excite.co.jp), Lycos 日本語版 (http://www.lycos.co.jp), Infoseek 日本語版 (http://www.infoseek.co.jp)。

クエリ群の候補として『現代用語の基礎知識 1999 年度版』(17737 語) から検索キーを抽出し、各検索サービスでの検索結果の URL 数が 50 ～ 400 の範囲に入った 1085 個を実験クエリとして採用した。これらクエリを検索サービスに与えて取得した URL 群のうち、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。各検索サービスごとのこれらクエリに関する無効URLの存在率を表1に示した。保有データ数を公開しているgooおよびLycosを基準にして、先に述べた推定方法を用いて日本語Webページ総数の推定を行った。

表 1: 各検索サービスにおける無効URL存在率

検索サービス	無効URL存在率	標準偏差
Excite	0.0926	0.0487
goo	0.0564	0.0457
Infoseek	0.0767	0.0407
Lycos	0.134	0.0526

5 結果とまとめ

現在 (1999年7月) の日本語Webページ総数は表2および表3のように推測される。表2ではS_aとして1700万ページを有していると称するgooを元に、表3ではS_aとして3000万ページを有していると称するLycos日本語版を元にして、対に選んだ検索サービスの結果から算出している。

表 2: gooを元にした日本語Webページ数の推定

対とした検索サービス	P_a	標準偏差	推定ページ数
Excite	0.334	0.150	5100万
Infoseek	0.321	0.144	5300万
Lycos	0.281	0.132	6050万

表 3: Lycosを元にした日本語Webページ数の推定

対とした検索サービス	P_a	標準偏差	推定ページ数
Excite	0.394	0.113	7610万
goo	0.357	0.123	8410万
Infoseek	0.450	0.126	6670万

いずれの結果についても、従来の研究 [3, 4] による値から急激な増加を示していることがわかる。この方法では、何処からもリンクされていない閉じたURL群や、アクセス制限されたページ、ロボットを規制しているページなどの非公開Webページを含まない。このために、これらの推定値は検索サービスが収集することの出来る範囲での大きさであり、実際の日本語Webページ数はこれよりも大きいと予測される。いずれの表においても、算出した値の信頼性を調べるために、各クエリから得られた結果の標準偏差も併せて示した。

表4には、各クエリごとに4つの検索サービスが返した全URL数に対する各検索サービスが所有するURL数を相対カバー率としてまとめた。表4をグラフ表示したものである。これら以外の検索サービスが所有するはずのURL数も考慮すると、これら4つの検索サービスが検索対象としてカバーしているWebページ数の割合はさらに低くなると考えられる。

表 4: 各検索サービスの蓄積URL数の比較

検索サービス	相対カバー率	標準偏差
Excite	0.315	0.0882
goo	0.368	0.127
Infoseek	0.400	0.116
Lycos	0.473	0.109

今回の研究では、日本語公開Webページ数を推定するために利用した検索サービスが公表している保有データ (URL) 数そのものがあいまいな数値であるため、推測したWebページ数自体に絶対的意味合いを持たせることはできないが、2つの推定値が似通っていることから推定値は日本語Webページ数の第一次近似を与えていると考えられる。

本研究による調査を定期的に行なうことで、日本語Webページの増加数、増加率などを推定することが可能である。また、この研究を応用して検索サービスの比較評価も可能である。

参考文献

[1] Steve Lawrence, C.Lee Giles, *Searching the World Wide Web*, SCIENCE 280, 99 (1998)

[2] Steve Lawrence, C.Lee Giles, *Accessibility of information on the web*, NATURE 400, 107 (1999)

[3] 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)

[4] 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)