2X－06

ドメイン毎のWebページ数の偏りを考慮した日本のWebページ数推定調査

西村真幸　山名早人
早稲田大学 理工学部 情報学科

1．はじめに

近年のインターネットの目覚しい普及により、Webページ数は急激な増加傾向を辿っている。しかし、実際にどのくらいのWebページ数が存在するのか、また現状の検索エンジンではどのくらいのWebページをカバーしているのか、といった正確なデータが現状ではあまりない。こういった理由から、Web技術の基礎研究として、Webページ数に関する調査方法の研究が重要になってきている。

Netcraft社[1]の調査によれば、2001年12月の時点での全世界のWebサーバ数は約3600万台であり、S.Lawrence氏の提案した手法[3]から比率を利用して試算すると、世界のWebページ数は約50億ページ存在することになる。しかし、現段階で最もカバー率が高いGoogle[2]でもカバーしているのは約30億ページである。Webページ数の増大は確実なものであり、巨大なWeb空間の実数調査はもはや不可能な状態である。

本稿では、S.Lawrence氏の推定サーバ数と1サーバあたりの平均Webページの積によるWebページ数推定方法[4]を基本的なモデルとした上で、jpドメインを日本のWebページと定義し、第二レベルのドメイン毎にWebページ数の偏りがあることを考慮した日本のWebページ数推定調査を考案した。

The Presumption of the Japanese Web Pages Based on the Deviation of the Number of Web Pages for Every Domain

Masayuki Nishimura, Hayato Yamana
Department of Science and Engineering, Waseda University

2．Webページ数推定調査の研究

この種の研究で有名なものとしては、S.Lawrence氏らが行ったWebページ数推定調査がある。S.Lawrence氏らは、検索エンジンを利用したWebページ数推定方法で1997年12月の時点で3億2000万[3]、推定サーバ数と1サーバあたりの平均Webページの積によるWebページ数推定方法で1999年2月の時点で8億のWebページが世界には存在するとした[4]。

また来住氏らは、日本の検索エンジンで収集したWebページを日本のWebページと定義し、日本の検索エンジンを利用したWebページ数推定方法で1999年11月の時点で1億2000万[5]、さらに2000年10月の時点で2億5600万のWebページが存在するとした[6]。

3．調査方法

2001年12月の時点で、約28万のjpドメインが登録されている[7]。そこから予約ドメインを除外すると、約26万の接続されているjpドメインが存在する。この接続されているjpドメインに対して、約5週間第二レベルドメイン毎にランダムに50ドメイン(汎用ドメインは1ドメイン、COドメインは100ドメイン)選択し、全体で約501ドメインを対象としたWebページ収集を行った。今回の調査では、ドメインの選択は全体の約0.2%を抽出した。

次に各第二レベルドメイン毎のWebページ数に偏りがあると考慮し、選択した各第二レベルドメイン毎にWebページが平均何ページあるかを計算した。そして、各第二レベルのドメイン数と平均ページを乗算し、最後に各第二レベルドメインのドメイン毎に求められたWebページ数を加算して、日本のWebページ数を推定した。

4．調査結果

調査結果を以下の表にまとめる。

表1　第二レベルドメイン毎の平均ページ数

(調査ドメイン数: CO100 ドメイン、その他 50 ドメイン)

ドメイン名	1ドメインあたりの平均ページ数	1ホストあたりの平均ページ数
AD	895	481
AC	4551	182
CO	1069	306
GO	3968	922
OR	338	282
NE	1118	299
GR	177	167
ED	236	207
地域	1449	717

表2　日本の Web ページ数推定結果

ドメイン名	第2レベルドメイン数	1ドメインあたりの平均ページ数	結果
AD	299	900	269100
AC	2679	4491	12031389
CO	214152	1200	256982400
GO	669	4111	2750259
OR	13706	356	4879336
NE	16931	1533	25955223
GR	10416	370	3853920
ED	3139	262	822418
地域	3854	1511	5823394
汎用	1	3348	3348
合計			313370787

表1より、各第二レベルドメイン毎の平均ページ数にはそれぞれ違いがあり、偏りがあることがわかる。S.Lawrence 氏の方法[4]では、選んだサーバ数とそこで得た Web ページ数から1サーバあたりの平均 Web ページ数を得ていた。表1より、今回考案した推定調査のほうが[4]の方法よりも細分化して調査しているので、より精度の高い Web ページ数の推定値を得ることができると考えられる。

また表2より、今回考案した方法によって、日本の Web ページ数は 2001 年 12 月の時点で約 3 億 1300 万ページ存在するという推定結果を得ることができた。

5．おわりに

本稿では、各第二レベルドメインの特徴を利用して日本の Web ページ数推定調査を行った。その結果、日本の Web ページ数は約3億1300万存在するという推定結果を得た。

今回は推定調査に約5週間費やした。しかし WWW は日々内容が更新され、WWW サーバが管理面からアドレスを変更することなどを考えると、WWW の研究に再現性を求めることは難しいが、できるだけ短期間に調査を行う必要がある。

今後、複数サーバ名の IP アドレスの割り当て、Ipv6 の導入などの影響により IP アドレス分布は大きく変化すると予想される。Web ページ数の推定調査には、さらに新しい方式の開発が重要になってくると考えられる。

謝辞　本研究の一部は、文部科学省科学研究費補助金奨励研究(A)（課題番号：13780255)及び早稲田大学特定課題研究(2001K-036)に よるものである。

参考文献

[1] Netcraft, http://www.netcraft.com/

[2] Google, http://www.google.com/

[3] Steve Lawrence, C. Lee Giles: "Searching the World Wide Web", Science, Vol.280, No.5360, Issue 3, pp.98-100 (1998.4)

[4] Steve Lawrence, C. Lee Giles: "Accessibility of Information on the Web", Nature, Vol.400, pp.107-109 (1999.7.8)

[5] 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)

[6] 来住伸子, 大森貴博, 水谷正大,小川貴英: "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", 情報処理学会論文誌,データベース, Vol.42, No.SIG 8, pp.47-55 (2000.10)

[7] -: "jp ドメインリスト", JPNIC (2001.12)