調査研究論文

インターネットコンテンツ統計に関する調査研究

通信経済研究部主任研究官　　中島　睦晴
研究官　　島田　博也

キーワード
web統計　サーチエンジン

[概要]

インターネットは90年代に入り、ワールドワイドウエブ（以下WWW）の普及によって、その社会的経済的重要性と影響度が飛躍的に増大しつづけている。このWWWが、我が国においてどの程度拡大もしくは変化し、それが何によってもたらされ、どのような波及効果を持つかを究明する事は、ｅ-Japan戦略を推進する行政サイドにとっても大変重要なテーマであると言える。

本稿では、郵政研究所での取り組みの成果として、日本のJPドメインの時系列的発展について諸データを元に報告するとともに、WWW規模調査研究の世の中におけるこれまでの流れを振り返りつつ、郵政研究所の取り組みを評価位置付けし、今後の取り組みへの展望を述べる。

1　インターネットとWebの関係

1-1　インターネットの歴史とWeb

インターネット自体の歴史は古く、1969年、米国のARPANETにおいてホストコンピュータ間のデータ通信が史上初めて成功した事に端を発するとされる。

しかし、インターネットが現在のような機能を発揮するのは最近の事で、核心となるWorld Wide Web（以下WWW）とブラウザの原理論が発表されたのが1989年の事である。

インターネットが世に提供してきた主たる機能によって、インターネットの歴史を整理すると図１の様になる。

こうして見ると、現在広くイメージされているインターネットとは、ここ10年ほどの事象である事が分かる。80年代まではメールによるテキストデータのやり取りが、インターネット機能の主幹であった。

1-2　現代インターネットの特徴

現在のインターネットを特徴付けている構成要素は大きく３つある。マルチメディアデータを取り扱い、HTML言語によって整理統合されたサーバ上の「サイト」、サイト同士をリンク構造により結びつける「WWW」、そしてこの２つを

図１　インターネットの変遷とWebの位置

軍事用インフラ時代 （70年代）組織ー組織型	破壊に強い分散ネットワークの実現（米国） 移動通信の条件整備（艦船・航空機の取り込み、TCP/IP） 72年、電子メール開始
学術用インフラ時代 （80年代）個人ー個人型	距離と時間の克服（電子メール・電子ニュース） 89年、WWWとブラウザの原理論発表
商用インフラ時代 （90年代）発信者ー閲覧者型	個人が利用できるコストと操作性の実現 特定、または不特定多数を対象とするコンテンツが主役 ●安価な接続サービスの普及 ●WWWの普及（サーバ・ブラウザ） ●検索エンジンの普及
個人のマルチメディア情報発信容易に？ （2000年代）	安価なレンタルサーバの普及 安価なドメイン取得管理代行サービスの普及 IPv6実用化、IPの広範な個人所有可能に PtoP技術の普及、サーバなしに個人が情報発信可能に

最大限に活用する為のツールである「サーチエンジン」である。

（※サイトはブラウザ機能の下位に整理し、WWWに含めるべきと言う議論もあり得ると思うが、情報所在の主体（コンテンツ）として、ここでは敢えてネットワークとは切り離して整理した。）

80年代に比べて、90年代以降のインターネットが発展した点は大きく２つある。ひとつは従来構想のみで普及していなかったオープンなネットワーク上でのマルチメディアの活用を現実のものとした事である。もう一つは利用者を世界レベルにおいて一般の市民や組織にまで拡大させた事である。この事は、社会において消費される情報量を飛躍的に増大させた。インターネットの効用について、誕生以来一貫しているのは、情報流通における距離と時間と費用の克服である。インターネットの主たる機能の時代的変遷は、この効用の適用範囲、普及範囲の拡大の歴史とも受け取れる。今日に至ってはその影響は、個人の生活から経済のあり方にまで及ぶ。

２　インターネットWeb調査の歴史と推移

２−１　インターネットWeb調査の意義

⑴　ｅ−Japan戦略に資する情報

今日政府が展開するｅ−Japan戦略は、このインターネットの効用の大きさを率直に評価し、これを社会、経済の活性化に活用しようとするものである。ｅ−Japan戦略は単年度のプロジェクトではなく、５年にわたる取り組みを期すものとなっている。

ネットワークのブロードバンド化は、その施策面での基幹のひとつであるが、ブロードバンドは絶対的基準ではなく、あくまで過去との比較における相対的な概念である。従って施策として見る場合は普及世帯数に加え「５年以内に能力的にどの程度までのブロードバンド化を後押しすべきなのか」という議論が必要であると考えられる。その為には「どの様なコンテンツがどの様に活用される必要があるか」、というビジョンを描く必要があるが、これには単年度の取り組みの成果をネット上において現れた変化として読み取って利用動向を把握し、これと利用者の意見を活用して次のステップを描く作業を進める事が望ましいと

考えられる。

⑵ Web統計の重要性

問題はネット上にあらわれる変化をいかに把握するかであるが、この為には、Webに関する時系列的な統計調査が必要となる。２つの指標が重要である。一つはブロードバンド化が活かせるコンテンツの普及状況、もうひとつはそのアクセス状況である。

コンテンツの普及状況については、質的側面と規模的側面がある。質的側面はコンテンツの内容に関する変化を追う事であり、テキストマイニング等を用いた言語処理的な分析と分類が有効であると考えられる。規模的側面はサイトの数とそのページ数、そしてそれらを構成するデータの種類（HTML、画像、動画、音声等）別の量など、インターネットの規模的な発展を追う事である。質的側面にしても、社会的経済的効用を発揮するには然るべき規模的発展が伴う事が前提となるので、コンテンツの普及状況をはかる軸となる指標として、規模的発展の度合いを確認する統計は非常に重要である。

従い、Webの規模的統計を活用する事は、e-Japan戦略にとっても大変重要であると言える。

２-２ Web調査研究の分類

⑴ 大きく２種類の調査

Webそのものが出現してまだ10年余りの存在であるから、その調査研究もまた大変新しい分野である。この10年間ほどの間に、主に２種類のタイプのWeb調査が登場している。ひとつはWebの規模に関する調査、もう一つはWebの構造に関する調査である（これに加え、最近では新分野として前述のテキストマイニングを用いた質的側面の研究が注目され始めている）。いずれも単発的な研究が中心で、時系列的に定期調査を行っている例は極めて少ない。

⑵ Webの規模に関する調査の分類

Webの規模に関する調査はそのアプローチにより主に３つの型が存在する。一つ目は複数のサーチエンジンの検索結果を活用し、そのアウトプットの違いから統計的手法を用いてWeb規模全体を推測しようというものである。二つ目はIPアドレスのサンプリング結果から推定するものである。そして三つ目は単独のサーチエンジンの走査結果をもとに線形近似によって推計する手法である。以下にそれぞれの内容について述べる。

２-３ 複数のサーチエンジンの組み合わせを活用する手法

⑴ 手法誕生の背景

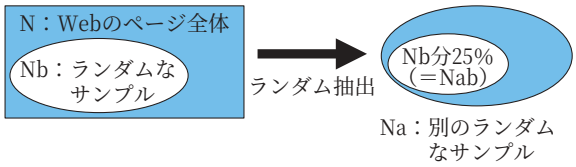
この手法の背景には勃興期にあったWeb検索エンジンの存在がある。多くの検索サイトが設立され、それぞれが「世界最大の検索規模」を謳い文句に激しい競争を繰り広げる中、この検索エンジンの実力を評価し、Web上の情報とは果たして十分に活用できる体制が出来上がっているのかどうかを検証するのがこの型の研究の問題意識である。従って必ずしもWebの正確な全体像を描く事にこだわらず、サーチエンジンの実力の測定、サーチエンジンの改良にむしろ主眼を置くものも多い。

⑵ 手法の仕組み

この手法の原理は以下の様なものである。今ここにWebの全ページの集合体Nがあるとし、その中にランダムにページを集めた集合体Nbがあるとする。Nからランダムにページをピックアップすると、それがNb内にヒットする確率は等しくNb/Nである（図２）。そうしてピックアップしたページを集めてNaという集合体を作ると、

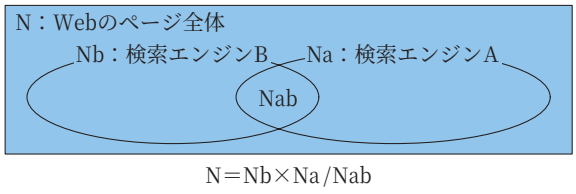
その内容もNb/Nの比率でNbにヒットするデータが混在しているはずである（図２）。と言う事は、今仮にNaの中でNbにも含まれるデータの割合が25%であるとするなら、Nb/N、つまりNの中でのNbのシェアが25%だという事になる。そこでNbの全数がわかっているなら、逆算して全Webページの数が計算できる。

図２　Naの作成とN数の推定



このNa、Nbの集合体の役割を検索エンジンにやらせようというのがこの手法である（図３）。

図３　複数の検索エンジンの組み合わせを活用したWeb測定手法の原理



サンプルになるキーワードの集合体をつくり、その検索から推計を行う。比較的手軽に調査が実施できる為、国内にもこの手法を用いた研究を行うグループが存在する。

代表的な研究にSteve LawrenceとC. Lee Gilesが1998年３月に発表した「Searching the World Wide Web」があり、これは６つの検索エンジンAltaVista, Excite, HotBot, Infoseek, Lycos, Northern Lightの検索カバー範囲を利用してインターネットの総ページ数を推計し、これら検索エンジンがどの程度全体をカバーできるのかを推定する目的でおこなわれたものである。

⑶　手法の問題点

Webページの全数調査をする上で、この手法には無視できない問題点が存在する。検索エンジンが保有するデータはWebページの無作為な集合体ではなく、作為的な集合体である。すなわち、「利用者に評価してもらう為には如何にWebページのデータを保持していなければならないか」、という問題意識の元に、事業として集められた情報の集合体である。従って利用者の登録がよくあるサイトには注意を払っているが、利用者に人気のないサイトには興味がないという方針のもとに集められている。。

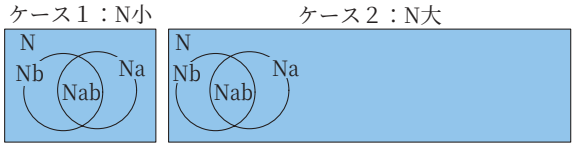
しかも検索エンジンは、データの収集にあたって主要な手段としてWebのリンクを辿るという手法を用いる。従ってリンクが密に張られているページはあらゆる検索エンジンに見つけてもらいやすく、リンクが薄いページはあらゆる検索エンジンからこぼれやすい。前出のLawrenceとGilesも1999年７月に発表した研究「Accessibility of Information on the Web」で「検索エンジンはサイトから見ると不平等である」としてこの問題を指摘している。

こうした事から先程の集合体のNabの収集に問題が生じる。NabはWebのリンクが発達してくれば自動的に膨張し、リンクが衰退していけば自動的に減少する。また利用者に人気があれば事業者のオペレーションによりNabに入れるが、人気がなくなればDBの更新時に消去されていく。つまり世間の話題の変遷で影響を受ける。そしてリンクがうすく、人気もないサイトは最初からNabとの縁が極めて薄いという事になる。すなわちランダムサンプリングによる生成物にはほど遠く、Nabはランダム抽出された場合と比較して、実際には大変大きな集合となっていると考えられる。これが無作為抽出と比較してどれだけの偏りがあるかを割り出せない限り、ここから統計的手法に

よってWebページ全体を推計する事には無理がある。

国内で消費される全ての日用品の種類数を推定するのに、コンビニチェーン同士の品揃えの違いからこれが割り出せるかを考えると、この手法をWebの規模調査に適用する際の問題点がよく理解できる。

尚、検索エンジン同士の活動の類似性からNabが数値的に大きくなる事について、Na/Nabが過小に算出され、ついてはNの推定が過小になるという議論もみられるが、Na、Nbともランダムサンプルでなく、Nab：Na＝Nb：Nが成立していない以上、この主張には根拠がない。すなわち、下記のケース１にもケース２にもなる可能性がある。



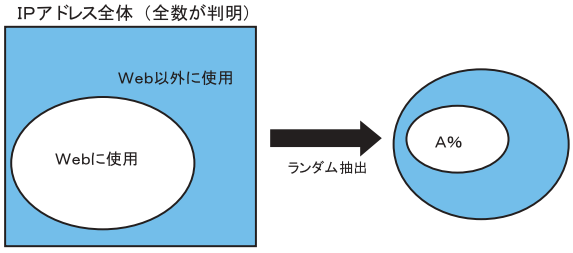
この手法はWeb規模調査の分野では大変ポピュラーな存在となっているが、こうした問題点から、郵政研究所はこの手法の採用を早い段階で断念し、独自手法の研究に着手した。この手法の価値は、全数調査よりもむしろ検索エンジンの機能を測定できる事にある。Nabを知る事でネットワークの中心的ページの実態が分かり、各エンジンのユニークな部分がどの程度広がりを見せるかによって検索エンジンの使い方を提言できる。またリンク構造の解明を通じてネットワークの構造の変化等を明らかにしていける。全数の把握はこの手法のテーマの一部に過ぎないと言える。

2－4　IPアドレスのサンプリングを活用する手法

(1)　手法の概要

これはインターネット上のノード（ネットワークに接続されているコンピュータやハブなどの機器）に対して必ず付与されるIPアドレスをサンプリングし、その中でWebサイトにヒットしたものの数値を元に、IP中のWebサイトの比率を割り出し、IPアドレスの総数から計算して推定しようという手法である。Webサイトの総数が分かれば、サイトの平均ページ数をサンプリングで推定し、これをかけ合わせる事でページ総数も推定できる（図４）。

図４　IPアドレスのサンプリングを活用する手法



IPアドレスは全数が判明しているので、Webに使用されたIPアドレスの割合が判れば、Webの全数が計算できる

IPアドレスは32bitのアドレス情報で、IPを用いて利用されるインターネット上の機器に必ず付与される事になっている。特に、Internetに接続される場合、世界レベルで唯一のIPアドレスを割り当てなければ、インターネットは成り立たない。このためにIANA（Internet Assigned Numbers Authority）という国際機構が存在し、アドレスをトラブル無く世界中に割り当てる任務を遂行している。実ユーザーに対する割り当て業務は、各国に国別組織（日本ならJPNIC）が存在し、これを遂行している。

IPアドレスはその全数が判明していてランダムサンプルが可能であり、複数の検索エンジンを活用する手法に比べると客観的な成果が期待できる。

代表的なものにOCLC　Research社が1997年に出した論文、「A Methodology for Sampling the World Wide Web（Edward T. O 'Neill、Patrick D. McClain、Brian F. Lavoie）」がある。また、

やはり前出のLawrenceとGilesも、1999年７月に発表した論文「Accessibility of Information on the Web」でこの手法を用いた調査結果を発表している。

⑵ 手法の問題点

この手法は従来は問題のない理想的な方法であったが、近年においては各種ネットワーク技術の進歩が障害となりつつある。すなわち一つのIPアドレスを複数のサーバが使える様にする技術の発達（IPマスカレード等）の発達により、あるIPアドレスの先にドメイン的に独立した複数のサーバが存在し、しかもいくつ存在するか調べる手立てがないという問題が生じている。IPアドレスを一つ取得する事が、いくつのサーバを取得した事になるのか判断できないとなると、信頼できるサーバ数やページ数の推定は不可能となる。

2-5 単独のサーチエンジンの走査結果をもとに推計する手法

⑴ 郵政研究所の取り組み

郵政研究所がシンクタンクのアライド・ブレインズ㈱と共同開発し、４年間実調査を実施してきたのがこの手法である。この手法により、半年に一度（２月、８月）日本の代表的ドメインであるJPドメインに関して、統計データを整備し発表してきた。

これは単独のサーチエンジンの走査結果が、時間を経るにしたがいリンク構造が均質化して、進捗グラフが直線を描くようになるのを利用し、最終的に総数を線形近似で推定する手法である。この手法はユニークなものであり、他に事例を見ない（特許出願中）。

尚、一部にこの統計データと、商用サーチエンジンの発表総ページ数を比較して乖離していると主張する論文があるが、商用サーチエンジンがJP以外の日本語サイト（Com、Netなど）の日本語ページも足し合わせているという基本的な事実を見逃しており、誤った主張である。

単独のサーチロボットを走査させる事でWebのページ数を求める研究としては、他にCyveillance社が2000年７月に発表した「Sizing the Internet」があり、独自のシステムを４ヶ月走査させる事で収集したリンク情報からWebの総ページ数を推計している。

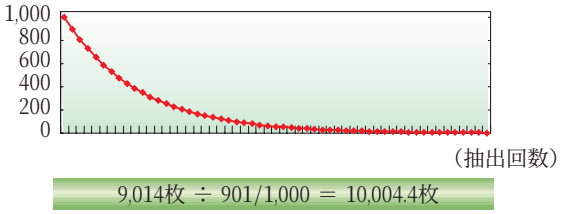
⑵ 郵政研究所の手法の原理

この手法の原理は以下のようなものである。今ここにカードの山があるとする。ここから全くデタラメに千枚抜き取り、裏に印をつけて山にもどす。山を良く混ぜ、もう一度千枚抜き取る。この中には最初に抜き取って印をつけたカードが混ざっている。印のついていないカードが何枚あるか数えた上で印をつけ、またこの千枚を山に戻す。この作業を繰り返し、千枚中印が付いていないカードが何枚ひけたかを都度数え、グラフを書いていく。

やがて一回で引ける印のないカードの数は序々に減り、グラフの傾きは０、すなわち横に寝た状態へ、線は曲線から直線へと限りなく近づいていく。千枚引いても印済のカードがほとんどを占める様になる。この時点で、千枚中に占める印のついたカードのシェアを計算し、過去延々と印をつけたカードの合計数をこのシェアで割る。そうすると山の正確な合計数が出る（図５）。

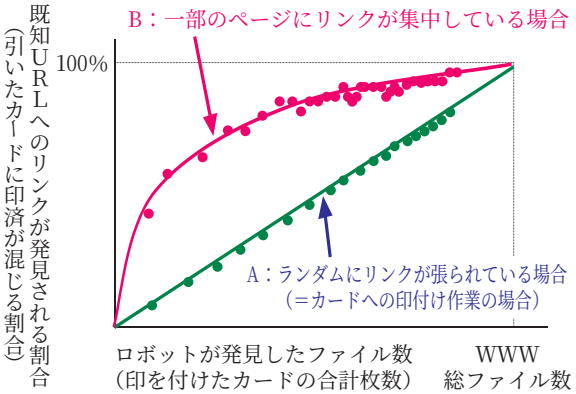
この作業を繰り返しつつ、横軸に引いたカードに印をつけた枚数を足していき、縦軸に引いたカード中印の付いたカードが混じっていた比率をとると、図６のグラフＡのように直線を描いていく。これは、均質なカードをランダムに抽出した場合、新規に見つかる無印カードの枚数の低下度合と印済みカード比率の進捗スピード鈍化度合が

図５　予測の原理モデル



１万枚のカードの山から1,000枚ずつ抽出、裏に印をつけて戻す事を繰り返す
23回目の抽出：1,000枚中901枚に印済、22回目までの印の累計9,014枚

図６　線形近似に使用するページ発見実績の推移



同じ速度で進む為である。

しかしこの手法で実際のWeb調査を実施すると、Ｂの様な線を描く。これは前述のカードモデルの場合、ひいたカードはそれぞれ同じく１回のカウントしか行わないが、Web調査でリンクの数を数える場合は、その数が数万リンクのページもあれば、１リンクのページもあって均一でなく、確率論的にリンクを多く持つページから出現する為である。従って最初のうちはたくさんのリンクを持つページが次々あらわれてカウント数値の伸びも速く、徐々にリンク構造が小型化していき、速度が鈍り、リンクが少ないページばかりになるので、「均質化」も進み、直線に近付いてくることになる。そこで直線になったと判断できるところで線形近似により、延長線上の総数を推計する。

⑶　郵政研究所手法の今後の課題

単独サーチエンジンの走査結果から線形近似する手法は、理論的には正しく問題の無いものである。しかし、走査結果のグラフが線形に至ったかどうかを実務上見極めるのが大変難しい。日々の走査結果は実際には微妙に差異があり、グラフは蛇行線を辿る為、直線となったのか、あるいは微妙な曲線をまだまとっているのか、判断は非常に困難である。微妙な曲線が残っていた場合は若干の誤差が生じる可能性がある。

また、Webの規模拡大に追いつきつつ動作の安定性を損なわない様な性能の強化も年々困難になりつつある。さらに近年ネットワークのトラフィックやサーバアクセスの著しい混雑が発生しており、データ収集に時間を要しすぎる等、実務面の課題は多いのが実情である。

課題解決の為には、より高性能のサーチエンジンを採用し、ネットワークの混雑やWeb拡大のハンディを克服して、線形と見なせる状況まで十分なグラフを描く事、そして、線形の判断が正しかったのかどうか、他の手法による調査でそれを検証する事が必要である。

3　郵政研究所の手法に基づく調査結果について

3-1　サーチロボットLokiによるJPドメインの調査結果

これまでの調査結果から、JPドメインの特徴について述べてみたい。

⑴　ここ１年間の全体的傾向

ページの伸び率、ファイルの伸び率、データ量の伸び率は、ピーク時に比べると緩やかではあるものの、現在も堅実に増加しつづけている（表１）。今年２月の調査ではサーバ数19.7万台（29.6%）、ページ数6555万ページ（7.4%）、ファイル数17,388万ファイル（13.9%）、データ

量5002GB（25.7%）（(　) 内は対前年比伸び)。

サーバ数の伸びのうち、2001年２月のデータには、プロバイダー等のサーバに間借りしていた為ドメイン的には存在が表に出ていなかったサイトが、最近のレンタルサーバの普及で多数独立した影響が含まれる。従いこの時点のコンテンツの本質的な増加はより緩やかであったと思われる。

表１　JPドメインの統計的推移

	サーバ数（台）	ページ数（万ページ）	ファイル数（万ページ）	データ量（GB）
1998年２月	36,000	1,020	1,891	305
	…	…	…	…
1999年２月	75,000	2,950	5,822	1,024
	108%	189.22%	207.88%	235.74%
2000年２月	95,000	4,250	9,627	2,214
	26.67%	44.07%	65.36%	116.21%
2001年２月	152,000	6,101	15,260	3,980
	60.00%	43.55%	58.51%	79.77%
2002年２月	197,000	6,555	17,388	5,002
	29.61%	7.44%	13.94%	25.68%

下段%は対前年比

図７は98年２月の数字を100として、その拡大を経年で比較したものである。サーバ数で約5.6倍、ページ数で約6.4倍、ファイル数で約9.2倍、データ量で約16.4倍に成長した事がわかる。

⑵　単位あたりのページとデータ量の変化

図８は、単位あたりの情報量の変化（１万ページの平均データ量とサーバ１台の平均ページ数）を見たものである。

プロバイダー系サーバに間借りしていたサイトは小型のものが多い為、その独立によってサーバの平均ページ数は過去２年間減少傾向が続いている。

一方、画像やPDFの増加により、単位ページ数あたりのデータ量は過去２年間増加し続けている。つまりJPドメインのページはどんどん重くなっている。

図７　JPドメインのコンテンツ量発達の推移（98年２月を100とした場合）

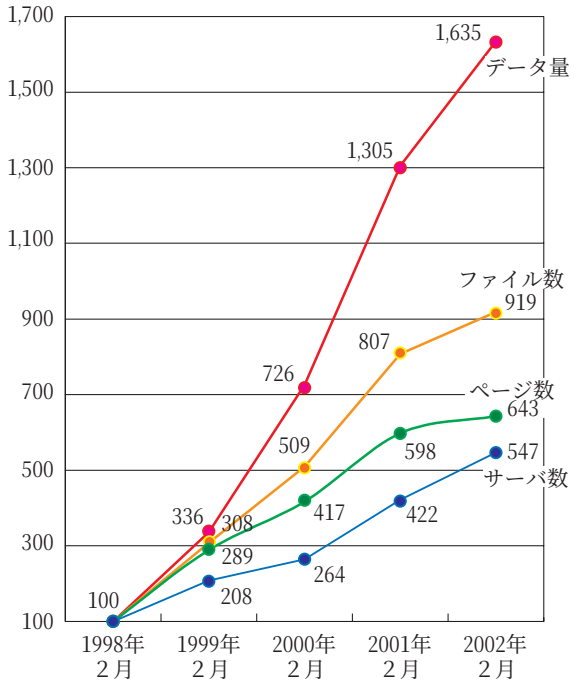
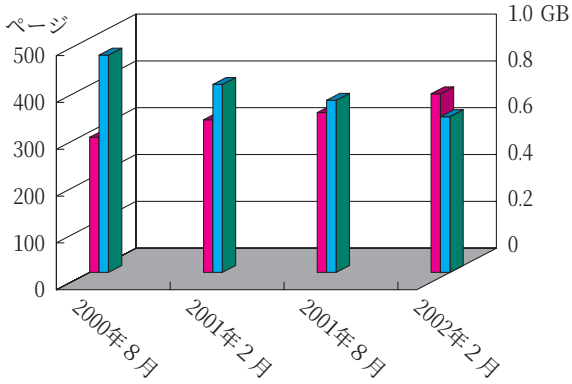


図８　JPドメインの平均データ量と平均ページ数の推移



調査時期	2000年８月	2001年２月	2001年８月	2002年２月
１万ページ当たり平均データ量（GB）	0.577	0.652	0.683	0.763
サーバ１台当たり平均ページ数（ページ）	464.2	401.4	367.6	332.7

⑶　JPドメインは画像データが基幹

ファイルの数で見ると、HTMLと画像が他のデータに比べ群を抜いて多い。中でも画像の増加はHTMLよりはるかに早く、年々両者の差は開いている（表２）。アップされているファイルの数から見るとJPドメインコンテンツの中心は画像であると言える。

表２　データ種類別ファイル数の推移

（万ファイル）

	1998年２月	1999年２月	2000年２月	2001年２月	2002年２月
HTML	1,023	2,953	4,255	6,107	6,558
画　　像	827	2,726	5,103	8,704	10,288
PDF	2	22	72	179	269
そ　の　他	38	121	196	270	273

⑷　PDFの増加について

過去４年間の調査の中で特に目立つのはPDFの増加である（表３）。全データ量に占めるシェアは98年２月で8GB、2.65%だったものが02年２月には1042GB、20.83%を占めるに至っている。この間データ量にして130倍の増加である。

理由としては

１）データ量が非常に軽い為、ナローバンドネットワークでも利用できるマルチメディアデータとして重宝がられている事、

２）過去の紙ベースの情報資産をネット化する上で、最も手間とコストがかからない方法である事、

３）PDFデータの閲覧ソフト（Adobe　Acrobat Reader）は無料で配布されていて急速に普及した事、

などが想定される。

PDF普及のデメリットとしては、現時点ではPDFデータそのものはわざわざHTML変換を行なわないとキーワード検索ができない為、ネットの情報活用という観点から、検索しづらいデータがネット上で増加している事になる。

この問題についてはAdobe Acrobat Readerの開発元であるAdobe Systems社が対応を検討している様である。尚、現在AcrobatがPDF形式をサポートする唯一のアプリケーションとなっている。

表３　PDFの占める割合の推移

	総　　数		P　D　F		構　成　比	
	ファイル	データ	ファイル	データ量	ファイル	データ
	［万Ｆ］	［GB］	［万Ｆ］	［GB］		
1998年２月	1,890	305	2	8	0.13%	2.65%
1998年８月	3,648	664	11	31	0.29%	4.61%
1999年２月	5,822	1,024	22	69	0.37%	6.69%
1999年８月	8,574	1,889	54	151	0.62%	8.02%
2000年２月	9,626	2,214	72	231	0.75%	10.44%
2000年８月	13,204	3,212	126	420	0.95%	13.09%
2001年２月	15,260	3,980	179	611	1.17%	15.35%
2001年８月	16,700	4,446	227	841	1.36%	18.92%
2002年２月	17,388	5,002	268	1,042	1.54%	20.83%

⑸ 動画・音声ファイルの推移とブロードバンドについて

動画・音声にはサーチロボットが収集できないファイルが多数ある為、部分的な推移しか把握する事ができず、この調査では全体的な動向は明らかでない。具体的に、以下の様なファイルが取得できない。

1）フラッシュのファイル
（一部とれるものがある）

2）ページに再生・停止ボタンが埋め込まれているようなもの

3）ストリーミングのファイル（存在は認識できるが、データ量を測定できない）
※ウインドウズメディアプレーヤーや、リアルプレーヤーのファイルが該当する

4）JAVA・Cgiにより再生されるもの

5）有料サイト、認証サイトのファイル
（無料の会員サイト等も含む）

6）リンク構造をもたないもの

4）5）6）は他の種類のファイルも事情は同じである。動画・音声に関しては3）4）は特に該当するものが多いと予想される。

以上から、現在サーチロボットにより取得できている動画・音声ファイルは、単純なリンク構造のみで使用できる環境になっているもの、という事ができる。これはサイト側から言えば、ビジネスへの関与等の特に強い目的性をもたない、言わば「参照ファイル」的なものと考えられる。

4　新たな調査手法の開発について

4-1　既存の各手法の特徴整理

従来の各手法の長所を活かし、または欠点を補い、第4の手法が構築できないであろうか。

調査手法は大きく2つの部分からなる。すなわち「代表性のあるサンプルとなるデータの収集」とそれをもとにした「正確な全体の推計」である。第2章の検討から、各手法はそのいずれか、または双方に問題があるという事になる。

今一度各手法の特徴を整理して見る。

1）複数のサーチエンジンを組み合わせる手法

・サンプル収集
サーチエンジンは恣意的なので、ランダムな収集にならない　代表性×

・全体推計
原理とする式自体が、検索エンジンに当てはまらない　推計信頼度×

2）IPアドレスのサンプリングを活用する手法

・サンプル収集
ネットワークの進歩が障害。代表性が確保できない　代表性×

・全体推計
サンプルさえ正確であれば極めて精度の高い推計が可　推計信頼度◎

3）単独のサーチエンジンの走査結果をもとに推計する手法

・サンプル収集
対象を限定すれば、ほぼ全数に近いサンプルを確保できる　代表性○

・全体推計
線形近似では見極め難しく、若干の誤差が生じる可能性　推計信頼度△

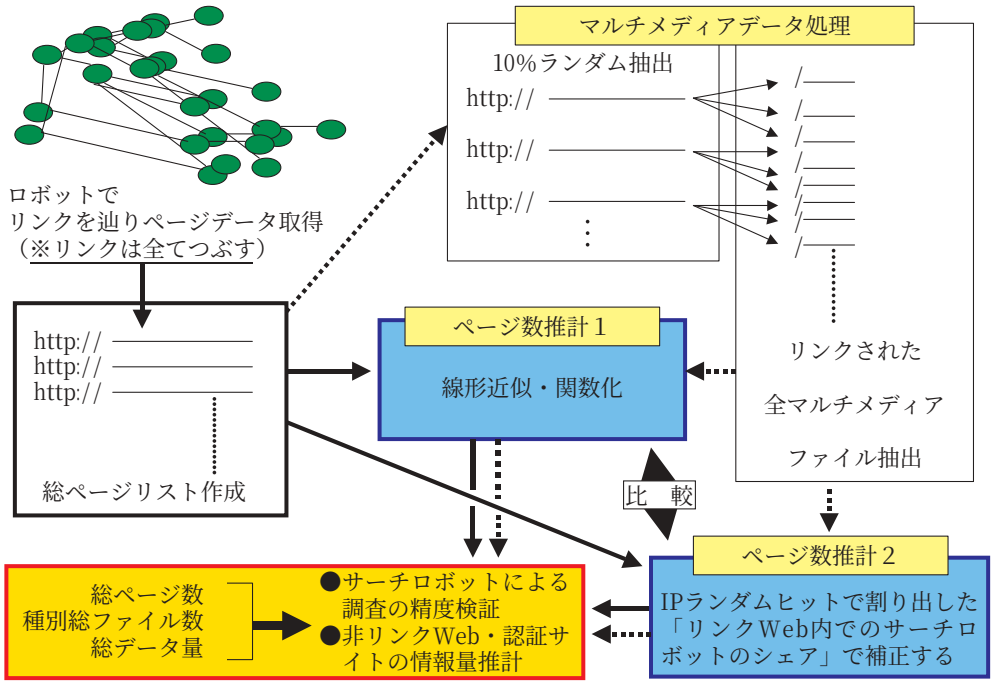
4-2　新たな手法の検討

⑴ 新たな手法の具体的手順

こうして見ると、単独サーチエンジンによるサンプルの収集と、IPアドレスのサンプリングを活用する全体推計の組み合わせが可能であれば、現状考えられる最も理想的な調査手法を開発できる事がわかる。

図９のような手法が成り立つ。これまでの郵政研究所の手法通り、まずサーチエンジンを活用して出来得る限りのリンクを辿り、ページを収集してリスト化する（図９では、「総ページリスト作成」の部分)。作成した総ページリストからランダムサンプリングにより、リンクされているマルチメディアファイルの数とデータ量を測る。一方、線形近似により先程の総ページリストのカバー率を推計する。この数値をもとにマルチメディアファイルの総数の推計値も割り出す。

図９　現在のWebを高精度調査する手法の概念



⑵　IPアドレスの用い方

次に国内の全てのIPアドレスを収集する。それをもとにrandom number generator（乱数発生器）により、サンプルIPアドレス群を複数作成する。サンプルIPアドレスのそれぞれの用途を調査確認する。そうすると、先程の総ページリストをサンプルIPアドレス群で打ち抜く様なイメージとなる（図10)。Webページに該当したIPアドレスのうち、作成した総ページリストに該当するもの（D)、リンクあるWebページだが総ページリストからこぼれたもの(E)、リンクがない、もしくはリンクの末端にあるWebページに該当したもの(A)をそれぞれ集計し、総ページリストの全体に占めるカバー率を割り出す。線形近似の推計結果が正しければ、総ページリストのページ数×$\frac{D+E}{D}$の値と一致する事になる。手法の精度を検証するために、両者を比較する。

この手法を実現させる事により、Web全体規模の推定について精度の問題はほぼ克服できるものと期待される。

図10　ページ数推計２の概念