

WWWの実態調査とその方法

安形輝(亜細亜大学) 石田栄美(慶應義塾大学大学院) 久野高志(作新学院大学女子短期大学部)

野末道子(鉄道総合技術研究所) 上田修一(慶應義塾大学文学部)

【抄録】 World Wide Web(以下 WWW)の調査はいくつか行われているが、WWW というメディアが抱える問題から適切な調査が行われているとはいえない。本研究では特にロボット収集の問題を解消するであろう、サンプリングの偏りを減らす工夫を提案している。実際にその手法を使って調査ファイル集合を作成し、ファイルの独立性に関する調査を行っている。

I. WWW調査研究における問題

WWWの調査を行うさいに問題となる主要な点は以下ようになる。

A. 全体の膨大な量

NEC 研究所の S. Lawrence らによれば1999年2月現在、約8億ページとされている¹⁾。さらに、彼らによる 1997 年12月の調査²⁾では約3億ページであったことから、未だに急速に成長し続けていることが推測される。このように WWW は膨大な規模かつ成長途中であるため、網羅的な調査を行うことは時間的に物理的に難しい。そのため、WWW の調査では原則的に一部のサンプルを何らかの手法で抽出し調査を行うことが必要になる。

B. 動的な性質

WWWは日々内容が更新されるものもあり、さらには WWW サーバが管理面からアドレス変更などがされることは日常的であるため、WWWは刻一刻と変化するものとなる。実際に、W. Koehler の研究³⁾では WWW の非常に短期的な性質を明らかにしている。このような性質のため、WWW の研究に再現性を求めることは難しく、また、できるだけ短期間に調査を行う必要が出てくる。

C. 1ファイル1単位の問題

WWW を扱う単位としては、サーチエンジン

を含めほとんどの研究で、1ファイルを1ページつまり1単位として扱っている。しかし、1ファイルは物理的な単位であり、内容的な面を考えると、妥当な単位かには疑問が残る。WWW において主題とファイルの関係を考慮すると、作者はある主題を1ファイルで表現することもある、逆に1ファイル中に複数の主題が表現することもあるはずである。従って、内容的な調査を行う場合、主題検索を行うサーチエンジンの場合、WWW の単位を検討する必要がある。

C. ロボット収集によるサンプリングの偏り

WWW の調査において調査対象を収集する最も一般的な手法は、サーチエンジンで使われているロボットを応用することである。ロボットはリンクを自動的に辿りながら、対象を拡大しながら再帰的にファイルを収集する。ただし、上述の WWW の膨大な量と動的な性質と収集側の物理的な制約(時間、回線速度、記憶容量など)のため、ロボットを動作させる場合、辿る回数を制限することになる。この制限の結果として、確率的によくリンクがされているファイルがより収集されやすくなるという問題を引き起こす。つまり、現在の WWW 実態調査を行っている研究のほとんどは、よくリンクがされているという性質を持つファイル群を中心とした偏った調査になってしまっている。

WWW の調査における問題のうち本研究では特に最後のロボット収集の問題を解消するような調査手法を考察する。具体的には、Lawrence らが行った研究のWWWサーバの調査に使われた手法を応用し、ロボット収集における問題点を解消するサンプル WWW ファイル群収集手法を提案する。さらに、WWW の単位を考える上で必要なファイルの独立性に関する調査も行う。

II. WWWの実態調査

今回の調査の流れは、最初に無作為なWWWサーバをインターネット上から調査し、それらに含まれる総てのファイル群から無作為なファイル群抽出を行うという二段階となる。さらに前半部には各 WWW サーバの調査も含まれるため、調査手法、さらに調査結果は全体として大きく3つに分けられる。単純にまとめると以下ようになる。

- 1)適当な量の重複なし無作為 IP アドレスを作成し、WWW サーバをチェックする(調査A)
- 2)各 WWW サーバ中を詳細に調査し、すべてのファイル群を分析する(調査B)
- 3)すべてのファイル群から抽出された内部リンクを無作為抽出し、ファイルの独立性について調査を行う(調査C)

A. インターネット上の WWW サーバ

1. 調査手法

この調査では無作為なIPアドレス群をチェックしWWWサーバを探索することを行う。

現在の標準である IPv4 では IP アドレスは 256^4 の 4 乗つまり概算すると約 43 億個存在している。しかし、すべてのアドレスが使われているわけではなく、また、使われているアドレスの一部が WWW サービスを提供するサーバとなっている。そこで、無作為に生成したアドレスを使い、インターネット上にどれほどの WWW サーバがあるかを推計することを行った。具体的に

は無作為な約 390 万の IP アドレス(先行研究は 360 万)の Web ポート(80 番)が開いているかをチェックした。実際には単にチェックするだけでは不用意な誤解を招くおそれがあるため、分散したシステムから、近傍アドレスの連続するチェックは避けることなどの工夫が必要となる。

2. 調査結果

調査は1999年 8 月 19～27 日にかけて実施した。結果として、390 万強調べた IP アドレスのうち、WWW ポートが開いているアドレスは 15917 であり、平均して 249 個のアドレスをチェックすることにより1つが WWW サービスを提供していることが明らかになった。

全調査IP	3965304
WWWサーバ	15917
認証あり	10727
認証なし	5190

表1 WWW サーバ数

B. 各 WWW サーバの調査

調査Aから得られたWWWサーバに対して、より詳細な分析を構造的な側面と簡単に内容的な側面の両面に対して調査を行っている。調査は 1999 年9月14～21日に行った。

1. 構造的な調査

この調査では独自のファイル解析システムを使い調査・分析を行った。対象は最初の 5000 サーバについてトップページのリンクを再帰的に辿り得られるファイル群である。ファイル中の HTML を解釈することで、リンク数、内部リンク数、外部リンク数、フレーム数、画像数等の項目について調査した。これらの結果を表2に示す。調査時期が調査 A とずれているため全 5000 サーバが読めたわけではなく、全数は 4318 サーバとなった。この表で左2列は全数が 4318 であり、その右2列は WWW 機器(読込バイト数が 100B 以下)と判断される分を除いた全数 3919 の結果となっている。1ファイルあたりのリ

	サーバ平均 (全:)	ファイル平均	サーバ平均 (全:)	ファイル平均	サーバ最高値
可読ファイル数	68.5		75.3		14726
不可読ファイル数	16.2		17.8		20024
リンク数	1360.4	19.9	1498.9	19.9	1072908
フレーム数	10.5	0.2	11.5	0.2	10193
画像数	532.1	7.8	586.3	7.8	166890
内部リンク数	1066.3	15.6	1174.9	15.6	962586
外部リンク数	300.7	4.4	331.4	4.4	157376
ファイル内リンク数	172.2	2.5	189.7	2.5	73927
HTMLリンク数	1256.5	18.4	1384.4	18.4	1024784
その他のリンク数	110.6	1.6	121.8	1.6	153704
読込バイト数	504841.0	7374.8	556232.1	7384.6	241695538

表2 各サーバの構造的なデータ

リンク数は 20 個弱、画像数は8個弱であることがわかる。

2. 内容的な調査

内容的な調査では人手により最初の 500 サーバについて言語、内容について調査した。ここでは原則としてトップページと必要な情報が得られるまで辿れるファイル群を実際にアクセスすることで調査している。言語として様々な言語を判定しているが、英語、日本語のように明らかに判断できるもの以外はドメイン名なども参考にして判定している。

これらの調査の結果は表3と表4にまとめられる。人手による内容的な調査では実際にアクセスした結果、アクセスできないサーバ、WWW 機器、システム付属の標準トップページのみのサーバを除いた 309 サーバが全数となる。

言語	サーバ数	%
英語	232	75.1%
ドイツ語	25	8.1%
日本語	11	3.6%
その他	41	13.3%

表 3 使用言語(全 309 サーバ)

各サーバでの主な(トップファイルから数ファイル辿るうちの)使用言語は、英語、ドイツ語、日本語、フランス語、中国語、オランダ語、イタリア語、スペイン語、ポルトガル語、カザフ語、ス

ウェーデン語、デンマーク語、ノルウェー語、フィンランド語、マレー語、ロシア語、朝鮮語であった。このうち日本語は英語、ドイツ語に次いで3番目によく使われている言語

となっている。

各サーバの内容のタイプは実際のファイルに表現されている概念から抽出し、309 サーバを見た時点で、企業、個人、製品、団体、店舗、プロジェクト、アダルト、大学、地方公共団体、学校、案内、協会、公共機関、雑誌、オンライン販売、ニュース、教会、事務所、図書館、サーチエンジン、システム、ホテル、メール、会議、学会、学区、本、その他のようになった。表3では内容タイプの上位10位までを示している。

タイプ	サーバ数	%
企業	177	57.3%
個人	20	6.5%
製品	18	5.8%
団体	13	4.2%
店舗	10	3.2%
プロジェクト	7	2.3%
アダルト	6	1.9%
大学	6	1.9%
地方公共団体	6	1.9%
その他	46	14.9%

表 4 内容タイプ(全 309 サーバ)

C. WWW ファイルの調査

調査 A、B の結果として得られた無作為な WWW サーバ中のすべてのファイル群から無作為に選択した 500 ファイルを対象としてファイルの独立性の調査を行った。対象ファイルを中心としてリンク関係にあるあるいは同じサブディレクトリ中のファイル群を手でチェックし、対

象ファイルを以下のカテゴリに分けている。

A: 1ファイル完結

表現される主題が該当ファイルで完結しているもの[例: 商品説明、リンク集(1主題)]

B: ファイル群の一部

ある主題が複数ファイルから表現され、そのファイル群の一部[例: 章分割された論文]

C: 1ファイル複数主題

複数の主題が1ファイル中に表現されているもの[例: FAQ 集、リンク集(複数主題)]

D: その他

英語、日本語以外のため解釈不能

解釈が困難な場合には該当ファイル中の主題検索を行い検索結果を出力することを仮定し、該当ファイル(→A)、ファイル群の代表(→B)、ファイル中の一部分のみ(→C)、というレベルを考えた場合、どのレベルで扱われるのが適当かという点から判断している。

調査結果を表5に示す。ファイル単位で扱うことができそうなファイルは全体の 30%弱に留まっている。

カ テ ゴ リ	ファイル数	割合
A: 1ファイル完結	141	28.2%
B: ファイル群の一部	232	46.4%
C: 1ファイル複数主題	39	7.8%
D: その他	88	17.6%
	500	100.0%

表5 ファイルの独立性

Ⅲ. 考察と今後の展開

A. 調査結果の考察

調査 A は先行研究と同じ形での調査となるため、比較ができる。調査 A におけるWWWサーバ数は少々増加しており、成長するWWWを裏付ける結果となっている。

調査 B では先行研究と一部重なる部分がある。例えば、ファイルあたりのバイト数の 7KB という値はほぼ同様のものとなっている。また、内

容のタイプ分けについては今回の方がより詳細なタイプ分けとなっているため単純な比較はできないが、アダルト関係サーバと識別された割合はほぼ同様となっている。一方で、言語に関しては先行研究では全く触れられていないが、今回の調査では英語以外が使われているWWWサーバがかなり存在するという結果となっている。

調査 C の結果では主題的にみた場合に1ファイルで完結しているものの方がファイル群の一部よりも少ないことが明らかであり、ファイル群の自動判定の必要性を示唆している。実際に 現在の多くのサーチエンジンはファイル単位で索引をしており、ファイル群の代表を登録しているディレクトリサービスがより精度が高いと言われる原因となっていることも考えられる。

B. 今後の展開

- 1) 無作為抽出されたファイル群の詳細な調査
- 2) ファイル群の質的な比較調査
- 3) ページ群の自動判定
- 4) サーバの規模的な差異の考慮

今回は考慮しなかったが WWW サーバごとの規模は大きく異なる⁴⁾ため、全ファイル群からの無作為抽出では特定の大規模サーバに偏る可能性が出てきてしまう。

【引用文献】

- 1) Lawrence, S.; Giles, C. L. "Searching the World Wide Web". Science". Vol.280, p.98-100(1998)
- 2) Lawrence, S.; Giles, C.L. "Accessibility of Information on the web". Vol.400, p.107-109(1999)
- 3) Koehler, W. "An Analysis of Web Page and Web Site Constancy and Performance". Journal of the American Society for Information Science. Vol.50, No2, p.162-180(1999)
- 4) Huberman, B.A.; Adamic, L.A. "Evolutionary Dynamics of the World Wide Web". <http://www.parc.xerox.com/istl/groups/iea/wwgrowth.html>

