日本の World Wide Web 情報空間: 1996 年 1 月のリンクデータ解析

中川 格, 石塚 英弘, 山本 毅雄
図書館情報大学
〒 305 茨城県 つくば市 春日 1-2
Tel: 0298-59-1111(代表), Fax: 0298-59-1093
E-mail:{itaru,ishizuka,yamamoto}@ulis.ac.jp
<URL: http://voyager.ulis.ac.jp/>

概要

ソフトウェアロボットが収集した 1996 年 1 月のリンク情報をもとに、国内の Web の情報空間の統計的調査を行なった。早稲田大学村岡研究室の田村健人から提供を受けたデータ (約 683 万リンク) を使用し、情報提供を行なっている組織数 (3,560) やサーバ数 (12,204) に関する基本的な統計量を算出した。全データのうち、活発に Web に取り組んでいる 1,235 組織 (active sites) 間の約 585 万リンクから、 1) これらの組織内および組織間のリンク数、 2) 参照率・被参照率を求めた。 active sites を属性により 6 グループに分類し、これらのグループの活動の性格を 1) リンク数の総数・平均値・中央値、 2) 平均参照率・相互参照率などから解析した。さらに 総合 Best 50 サイトを選出し、それらの関係を明らかにした。

キーワード

World Wide Web, 情報空間, 統計調査, Active Sites, 日本

Japanese Web Information Network: January 1996

Itaru NAKAGAWA, Hidehiro ISHIZUKA, Takeo YAMAMOTO
University of Library and Information Science
1-2 Kasuga, Tsukuba, Ibaraki 305, Japan
Tel: +81-298-59-1111, Fax: +81-298-59-1093
E-mail:{itaru,ishizuka,yamamoto}@ulis.ac.jp

ABSTRACT

Configuration and activity of Japanese World Wide Web (Web) information network were statistically studied using 6.83 million link data collected by a software robot (by Kento Tamura of Waseda University) in January 1996. Among 3,560 sites running 12,204 servers, 1,235 were chosen as active sites: links among them were 5.85 million, or 85.6 % of the total. For each active site, 1) the number of internal, external outgoing, and external incoming links, and 2) referring, and being refereed percentages were calculated. For six groups of active sites, 1) the total, the average, and the median of the number of links between groups, and 2) the average referring percentage, and bi-reference percentage were obtained. The above data were used 1) to analyze the activities of those groups, 2) to choose 50 most active sites, and 3) to elucidate their interdependence.

Keywords

World Wide Web, Information Network, Statistical Analysis, Active Sites, Japan

1. はじめに

本研究では、World Wide Web(Web) 上で情報提供を行っている組織間の参照関係を統計的に調査し、Web の情報空間の基本的な統計量を算出するとともに、その情報空間がどのように構成されているかを解析した。解析データはソフトウェアロボットが収集したリンク情報を用いた。

近年 Web は目覚ましく発展し、多くの情報が Web 上で提供されている。また、情報量の増加にともない、リンクにより形成される情報空間も複雑さを増している。このため、手探りでリンクをたどって求める情報を見つけ出すことが難しくなった。この問題を解決すべく、ソフトウェアロボットを利用した情報資源の探索 (information discovery) の研究が盛んに行われている。探索時に見つけられた情報は、所在情報データベースとして広く利用されているが、それらに関する詳しい統計量は報告されていない。

本研究の調査では、情報の内容には立ち入らずに、リンク情報のみを使用して国内の組織 ("jp" ドメインの組織) の参照関係を分析し、個々のページの更新に対してもある程度頑強な Web の情報空間の全体的な特徴を抽出する。このような情報空間の統計的調査は、現在の Web の状況を知るだけではなく、これまでの発展の軌跡を記録するとともに今後を予測する上でも重要であり、Web 上での情報検索を支援するための基礎データを与えるものでもある。

2. 解析に用いたデータ

本研究で用いたデータは、早稲田大学の田村健人が作成したソフトウェアロボット「千里眼」が収集した生データをもらい受けたものである。田村は 1994 年 12 月から千里眼を用いて国内のページ情報を収集し、所在情報データベースを作成している。本研究では、千里眼が 1996 年 1 月に収集したデータを解析した。オリジナルデータ中には、http オブジェクト、ftp オブジェクト、gopher オブジェクト、nntp オブジェクトへのリンクなど、様々なリソースへのリンク情報がある。ここで、http オブジェクトとは httpd (HyperText Transfer Protocol Daemon) が扱えるオブジェクトの総体である。その他のオブジェクトについても同様である。また、http オブジェクト間リンクの中には特殊なリンクとして、Delegate などの中継サーバを介したリンク情報が含まれている。

3. 分析結果

3.1 データのサンプリングと基本統計量

本研究では国内にある httpd オブジェクトとその間のリンクによって構成される情報空間を調査するために、以下に示すサンプリングを行なった。

1. 中継サーバを経由したリンクは、中継サーバを除き、リンク元からリンク先に直接リンクが張られている形に変更する。
2. httpd オブジェクト間リンク以外の場合はそのリンクを削除する。
3. 国外 (JP ドメイン以外) の httpd オブジェクトがリンク先あるいはリンク元になっているものを削除する。
4. httpd オブジェクトを提供している組織が実在しない場合はそのリンクを削除する。

このサンプリングにより、オリジナルデータに含まれる 6,829,256 リンクのうち、5,950,558 リンクが国内 http オブジェクト間リンクとして残った。これはオリジナルデータの 87% にあたる。本研究では、Web 上で情報提供を行なっている組織を、JPNIC の定めた「JP ドメイン名の割り当てについて」[3] と「JP ドメイン名 (地域型) 割り当てについて」[4] に基づいて、第 2 レベルドメイン名をもとに 6 つのグループに分類した。すなわち、ac.jp, ad.jp, co.jp, go.jp, or.jp と、その他の "others" である (表 12)。国内 httpd オブジェクト間リンクから、登録ドメイン名 (各組織のインターネット上での名前) とサーバ名を抽出し、グループごとにそれらの数を集計した。さらに、それらがインターネットに接続しているサイト数に対し、どのぐらいの割合かを調べた。その結果を表 1に示す。この表から以下のことがわかる。なおサイトとは、インターネット上で使わ

れているJPドメインの登録ドメイン名を所有している者(組織)とし、実際には「日本ドメイン名一覧表」[5]のエントリに対応するものとした。

- 全サーバ数の59%を教育および学術機関(ac.jp;以下「大学」と略す)のサーバが占めており、2位の企業(co.jp)とあわせると全体の約84%をこの2つのグループが占める。
- 全サイトの61%を企業が占めており、大学は16%で2位だった。
- 大学は一つのサイトで平均10以上のサーバを運営しているが、多くの企業は1サイトあたり1つのサーバを立ち上げている。このことから、企業では全社的なコントロールがおこなわれていると予測できる。
- インターネットに接続している全サイト(6,059)のうち、58.8%(3,560)がWeb上で情報提供を行なっている。
- インターネットに接続している全サイトに対するWeb上で情報提供を行なっているサイトの割合は、大学、政府関連機関、ネットワーク管理組織が非常に高く、いずれも80%以上である。

著者らはこれまでの研究[2]において、Webサーバは3種類のリンクの数により特徴づけられることを示した。それらはInternal Links、External Outgoing Links、External Incoming Linksである。サイトを単位とした場合にも同様のことが言える。その意味は表2のようになる。

上記3560サイト中には、Int. Linksの数が少なく、他のサイトからの参照(Ext. Inc. Links)数も非常に少ないサイトも数多く含まれている。本研究では、Int. Linksが10未満のものとExt. Inc. Linksが10未満の微小サイトを解析対象外とし、これらのサイトが提供するhttpオブジェクトがリンク元あるいはリンク先になっているリンクを国内httpオブジェクト間リンクから除いたものを解析対象リンク(サンプルセット)とした。この結果、1235サイト(active site)が解析の対象となり、サンプルセットに含まれるリンク数は5,845,417になった。表3に、微小サイトを含むサイト数とactive site数のグループごとの内訳と、全active siteに対する各グループのactive siteの割合を示す。この表より以下のことがわかる。

- 全体の約1/3がactive siteである。
- 企業サイトは全体の75%が微小サイトであるが、全active siteの約40%以上を企業が占めている。
- 団体(or.jp)とネットワーク管理組織(ad.jp)は絶対数は少ないものの、60%近くがactive siteである。

各グループごとに3種類のリンクの総数を表4に示す。この表から以下のことが読み取れる。

- すべての種類において大学が最も大きいシェアを占めている。
- ネットワーク管理組織はInt. Linksの数とExt. Out. Linksの数がほぼ同数である。
- 企業はInt. Linksの数にくらべ、Ext. Out. Linksの数が非常に少ない。

3.2 リンク数行列と参照行列を用いた分析

Web情報空間のより詳しい解析を行うためにリンク数行列と参照行列を以下のように定義した。

$$\text{リンク数行列 } L_{ij} \leftarrow \text{サイト } S_i \text{ からサイト } S_j \text{ へのリンク数}$$

$$\text{参照行列 } R_{ij} \leftarrow \begin{cases} 1 & S_i \text{ から } S_j \text{ へリンクがあるとき} \\ 0 & \text{ないとき} \end{cases}$$

リンク数行列の意義は自明であるが、参照行列を用いることの意義についてはこの節の後半で議論する。
先に述べた3種類のリンクの数はリンク数行列 (L_{ij}) を用いると式 1、式 2、式 3のようになる。

$$Int.\ Links\ 数 \quad = \quad L_{ii} \tag{1}$$

$$Ext.\ Out.\ Links\ 数 \quad = \quad \sum_{j \neq i} L_{ij} \left(= \sum_{j} L_{ij} - L_{ii} \right) \tag{2}$$

$$Ext.\ Inc.\ Links\ 数 \quad = \quad \sum_{j \neq i} L_{ji} \left(= \sum_{j} L_{ji} - L_{ii} \right) \tag{3}$$

各サイトを、これらの3つの量をもとにプロットしたものを図1に示す。この図からは、Int. Links と Ext. Out. Links が多い (右上) と Ext. Inc. Links も多く (直径大) なる傾向にあるが、Int. Links と Ext. Out. Links がそれほど多くなくても Ext. Inc. Links が多いものもあることがわかる。

あるグループに属するサイトが同じグループ内のサイトに張るリンクの量と、他グループのサイトに張るリンクの量を比較し、グループ間の参照関係にどのような差があるかを調べた。グループ間の参照量をグループ k からグループ l へのクロスリファレンス数 (N_{kl}) とし、式 4と定義した。

$$\text{クロスリファレンス数}(N_{kl}) \quad = \quad \begin{cases} \displaystyle\sum_{i \in G_k} \sum_{j \in G_l} L_{ij} & if\ k \neq l \\ \displaystyle\sum_{i \in G_k} \sum_{\substack{j \in G_k \\ j \neq i}} L_{ij} & otherwise. \end{cases} \tag{4}$$

ここで $k, l \in$ { ac.jp, ad.jp, co.jp, go.jp, or.jp, others} であり、 G_k はグループ k に属するサイトの集合である。各グループ間の N_{kl} の値を表5に示す。この結果から、ネットワーク管理組織グループ (ad.jp) 以外のグループは大学からの参照数が他のグループからの参照数よりかなり多いことがわかる。

グループごとの Int. Links の平均と、クロスリファレンスの平均は式 5および式 6となる。

$$Int.\ Links\ 平均 \quad = \quad \frac{\displaystyle\sum_{i \in G_k} L_{ii}}{g_k} \tag{5}$$

$$クロスリファレンス\ 平均 \quad = \quad \frac{N_{kl}}{g_k} \tag{6}$$

ここで $k \in$ { ac.jp, ad.jp, co.jp, go.jp, or.jp, others} であり、 G_k はグループ k に属するサイトの集合である。また、g_k は G_k に属するサイト数である。この計算結果を表6に示す。 表6からは以下のことがわかる。

- すべてのグループにおいて、Int. Links が各グループへの Ext. Out. Links よりはるかに多い。
- ネットワーク管理組織の Ext. Out. Links の合計と Int. Links の合計はほぼ等しい。
- 政府関連機関と “others” を除いて、それぞれのグループ内でのクロスレファレンスが多い。
- 企業は他グループへほとんどリンクを張らない傾向にある。

表6のような平均値は例外的に巨大なサイトに影響を大きく受け、代表的なサイトの特徴が現れていない可能性がある。そこで、3種類のリンクの各グループごとの中央値を調べた (表7)。表7と表6を比較するとから以下のことがわかる。

- 企業グループの代表的なサイトは Int. Links が 400、 Ext. Inc. Links が 100 に対し、 Ext. Out. Links が 4 と非常に少ない。

- 大学の代表的なサイトは Int. Links が 1200 で最も多く、他サイトの情報への参照数も、他サイトからの参照も多い。
- 各グループの Int. Links の平均値 (表 6) と中央値を比較すると、内部に大量のリソースを持ついくつかのサイトの Int. Links の量が平均値に大きく影響している。特に政府関連機関のグループにおいてその影響が顕著である。

これらのことからわかるように、小中規模のサイトが大半を占めているにも関わらず、リンク数による解析は、少数の巨大サイトに大きく影響されている傾向にある。そこで、参照行列を用いて割合による解析を行なった。以下に解析結果を示す。

まず、各サイトがどのぐらい多くのサイトを参照しているか (参照率; 式 7)、あるいは参照されているか (被参照率; 式 8) を調べた。

$$\text{参照率 } (RP_i) = \frac{100 \times \sum_{j \neq i} R_{ij}}{N_a} \tag{7}$$

$$\text{被参照率 } (REP_i) = \frac{100 \times \sum_{j \neq i} R_{ji}}{N_a} \tag{8}$$

ここで N_a は active site 数 (1235) である。図 2に各サイトの参照率、被参照率による散布図を示す。この図で X 軸が参照率、Y 軸が被参照率である。この図において大半のサイトが $\{0 \leq X \leq 10, 0 \leq Y \leq 10\}$ の範囲にあり、参照率・被参照率がともに高いものは重要なハブ・サイト (Hub Site) としての役割を持つと言える。

次に、同じグループ内への参照率と他グループへの参照率を比較し、グループ間の参照率にどのような差があるかを調べた。グループ k から l への平均参照率 (P_{kl}) は、式 9となる。

$$\text{平均参照率}(P_{kl}) = \begin{cases} \dfrac{100 \times \sum_{i \in G_k} \sum_{j \in G_l} R_{ij}}{g_k \times g_l} & if\ k \neq l \\ \dfrac{100 \times \sum_{i \in G_k} \sum_{\substack{j \in G_k \\ i \neq j}} R_{ij}}{g_k \times (g_k - 1)} & otherwise. \end{cases} \tag{9}$$

ここで $k, l \in$ { ac.jp, ad.jp, co.jp, go.jp, or.jp, others} であり、G_k はグループ k に属するサイトの集合である。また、g_k は G_k に属するサイト数である。平均参照率 (P_{kl}) の値は表 8のようになった。この表からは以下のことが読みとれる。

- ac.jp–ac.jp 間と go.jp–go.jp 間の平均参照率が他に比べ大きい。
- 異なる 2 つのグループ間の平均参照率は、1% – 3% 台のものが多い。

活性な参照関係 $(R_{ij} = 1)$ のうち、お互いに参照しあっているもの $(R_{ij} = R_{ji} = 1)$ がどれぐらいの割合かを調べ、国内全体の参照関係の歪みがどのぐらいあるのかを調べた。この歪みを調べるために、まず参照行列を 3 つの行列に分解した。参照行列 R は、異なる 2 つのサイト間の相互参照を示す行列と、片方からのみの参照を示す行列と、内部参照を示す対角行列に分解可能である。つまり、参照行列 R は双方向参照行列 R^b、単方向参照行列 R^s、対角行列 D(式 10) を用いて式 11のように分解できる。

$$
\begin{aligned}
R^b_{ij} &= \begin{cases} min(R_{ij}, R_{ji}) & if\ i \neq j \\ 0 & otherwise. \end{cases} \\
R^s_{ij} &= \begin{cases} R_{ij} - R^b_{ji} & if\ i \neq j \\ 0 & otherwise. \end{cases} \\
D_{ij} &= \begin{cases} 0 & if\ i \neq j \\ R_{ij} & otherwise. \end{cases}
\end{aligned} \tag{10}
$$

$$
R = R^b + R^s + D \tag{11}
$$

この分解 (式 11) により、各サイトの参照関係が非対称であることが明らかになった。これらのサイト間の参照関係の非対称性の度合を調べた。ある一つのサイト $Site_i$ と、$Site_i$ と参照関係あるいは被参照関係を持つ (複数の) サイトとの間の相互参照率を双方向参照行列 R^b、単方向参照行列 R^s を用いて式 12と定義した。

$$
i\text{番目のサイトの相互参照率} = \frac{100 \times \sum_j R^b_{ij}}{\sum_j (R^s_{ij} + R^s_{ji}) + \sum_j R^b_{ij}} \tag{12}
$$

さらに全体の相互参照率は式 13となる。

$$
\text{全体の相互参照率} = \frac{100 \times \sum_i \sum_j R^b_{ij}}{\left(\sum_i \sum_j (R^s_{ij} + R^s_{ji}) + \sum_i \sum_j R^b_{ij} \right)} \tag{13}
$$

この値は 13.9 となった。この結果、あるサイト i から別のサイト j への参照があっても多くの場合 j から i への参照は無いということがわかる。全体の相互参照率が低かったため、個々のグループ間の対称性を調査した。2 つのグループ k と l の間の相互参照率 (C_{kl}) を、式 13を拡張して式 14と定義した。

$$
\text{相互参照率}(C_{kl}) = \frac{100 \times \sum_{i \in G_k} \sum_{j \in G_l} R^b_{ij}}{\left(\sum_{i \in G_k} \sum_{j \in G_l} (R^s_{ij} + R^s_{ji}) + \sum_{i \in G_k} \sum_{j \in G_l} R^b_{ij} \right)} \tag{14}
$$

ここで $k, l \in$ { ac.jp, ad.jp, co.jp, go.jp, or.jp, others} であり、G_k はグループ k に属するサイトの集合である。C_{kl} の計算結果表 9に示す。この結果から、以下のことがわかる。

- ac.jp 同士、go.jp 同士の相互参照率がもっとも高く 30% ちかい。
- その他のグループは、“others” との間で相互参照率が高い。
- Web 上の参照関係の大半が片方向の参照であり、双方がお互いに参照しあっているものは少ない。

3.3 データ特性

ここでは、リンク数行列 (L) と参照行列 (R) の 2 つの行列からわかる、Web の 全体的なアクティビティについて議論する。

2 つの行列の特徴から、Web の全体的な特徴として以下のことがわかった。

- $R \neq R^T$ である。

  ⇒ Web のアクティビティを表す行列は非対称行列である。全体の相互参照率が 13.9% であることから、片方向の参照の方がかなり多いことがわかる。

- L の対角要素の合計は、全体のリンク数の 88% を占める。

  ⇒ Web 上のリンクのほとんどは、同一サイト内に向いており、他のサイトが提供する情報への参照はあまりない。

- R の全要素の 95.6% が 0 である。

  ⇒ 全体的な傾向として、あるサイトは少数の特定のサイトにリンクを張っている。

3.4 Best 50 サイトの分析

総合的に最も活発に Web に取り組んでいるサイトのアクティビティと、その間の参照関係がどのようになっているかを調べた。まず、総合 Best 50 サイトを選ぶために、5つの要素から各サイトのスコアを求めた。その計算式は式 15とした。

$$\begin{aligned} Score &= 1 \times Log_2(Int) + \tfrac{1}{4} \times Log_2(Out) + 1 \times (3 \times Log_2(RP)) \\ &\quad +1 \times Log_2(Inc) + \tfrac{3}{2} \times (\tfrac{10}{3} \times Log_2(REP)) \\ Log_2(x) &= \begin{cases} log_2(x) & if\ x \geq 1 \\ 0 & otherwise. \end{cases} \end{aligned} \tag{15}$$

式 15の作成にあたって、まずそれぞれの値を等級に分けるために log_2 スケールを使用した。それぞれの重みは Int. Links(Int) を基準とし、以下のような判断により決定した。

Out (Ext. Out. Links): Ext. Out. Links は他の人の財産を自分の財産とすべく採り入れたもので、謂わば他人の褌での勝負である。このため、Int. Links より重みを減らし、 1/4 にした。

RP(参照率): Yellow Page を提供しているようなサイトは、情報検索の際に有用であり Int. Links と同じ値にした。$Log_2(RP)$ を 3 倍しているのは $Log_2(Int)$ の値と最大値を揃えるためである。

Inc (Ext. Inc. Links): 他のサイトから参照されているということは、何らかの意味で重要であることを示しているため、Int. Links と同じ値にした。

REP(被参照率): 多くのサイトから参照されているということは、非常に重要であることを示している。このため、Int. Links より重い 3/2 にした。$Log_2(REP)$ を 10/3 倍しているのは $Log_2(Int)$ の値と最大値を揃えるためである。

このスコアの 上位 50 サイトを Best 50 サイトとし、その結果を表 10に示す。 この Best 50 サイトのアクティビティを表 11に示す。表 10と表 11の 2 つから以下のことがわかる。

- Best 50 サイトには、大学・プロバイダ・情報関係を扱う企業・政府の研究所が多く含まれる。
- Best 50 サイトのうちの半数以上 (32) を大学が占めている。
- Best 50 サイト間リンク (Int. Links も含む) が全 active site 間のリンクの半分近くを占める。
- Best 50 サイトはお互いにほとんどのサイトを参照している。

Best 50 サイト間のリンク数のシェア率は Int. Links を差し引いて計算しても全 active site 間の Ext. Out. Links の 28.7 % を占めていた。
Best 50 サイト間のアクティビティをわかりやすく表現するためにそれらの間の参照関係図を作成した (図 3)。

表11からわかるように、すべての参照関係を図に描き込むと、複雑で見にくい図になる。そこで閾値を設け、それ以上の参照数のものを描き込むという形式をとった。

まず、Best 50サイト間で、サイト S_i から サイト S_j へのリンク数が500以上のものを抽出した。この結果、31サイト間の 60個の参照関係が該当した。これらの関係は、地図中に太い方向付エッジで記した。次に、地図中に 50 サイトすべてを含ませるために、図に現れなかった 19 サイトに関してその最大の参照先を調べ図に描き加えた。なお、図中の各サイトを表す円の大きさはスコアをもとに 5 点刻に 5 段階に分けている。この図からは、以下のことがわかる。

- Webの参照関係は、学術的な部分 (大学や研究所を中心としたもの) と非学術的な部分 (企業やプロバイダを中心としたアクティビティ) とに大きくわかれるが、境界ははっきりしていない。
- 学術的な部分では、東大が巨大なハブサイトとしての役割を担っている。
- 非学術的な部分では、リムネットをはじめとするプロバイダが中規模のハブの役割を担っている。

図3は、 Webのアクティビティを的確に表しているが、複雑で多少見ずらい。そこで、サイト S_i が最も多く参照している外部サイト S_j への方向付エッジのみを記したものを図4に示す。

4. 考察

本研究では、日本国内のWebの情報空間の様々な統計量を算出するとともに、その全体的な特徴の解析を行なった。これはWoodruffら [7] が必要であると指摘したStructual Network Analysis のWebへの応用を試みた研究ともみなせる。

まず国内のWeb上で情報提供を行なっているサイト数やサーバ数などの基本的な統計量を算出した。これらのサイトのうち、活発に情報提供を行なっている1235 のactive siteについて、 1)active site 間のリンク数、2) 参照率、被参照率を調査した。さらにactive siteを6つのグループにわけ、グループ間のアクティビティについて、 1) グループごとのリンク数の平均値と中央値、2) グループ間平均参照率、3) グループ間相互参照率等についての解析を行なった。さらに、これらの量をもとに 総合Best 50サイトを選び、それらの間のアクティビティを調査した。

これらの解析を通してわかったことには、例えば、1) 各サイト内部への参照を意味するInt. Linksの合計が国内全体のリンク総数の大半を占めていること、2) Web上の参照関係は対称ではなくかなり偏りがあること、3) Webの情報空間には大学や政府の研究所を中心とする学術的な部分と企業や各種団体を中心とする非学術的な部分とがあるが、境界ははっきりしていなこと、などがある。これらの解析を通して得られた知識は、データ収集後ほぼ1年経った現在でも情報検索の際に十分に役立っている。これはWebの急激な発展を考えると驚異的なことである。このことは本研究が目標とした、多少の変化に対してもある程度頑強なWebの情報空間の全体的な特徴を発見することができたことを意味している。本研究で行ったようなWebの情報空間の調査を世界の各地域を対象として行ない、さらに各地域間の連関を解析することにより、真にWebの World-Wide な特徴を浮き彫りにすることができるだろう。

解析を通して得られた知識の中には、1) 参照率95%の日本科学技術情報センタ (JICST) はインターネット上の簡易団体名鑑として有用であること (注1)、2)　多くのExt. Out. Linksを持つORIONSは、情報検索の際に有用であること (注2)、3) 歴史的に重要な立場にあるNTTは多くのサイトから参照されており、依然として重要な立場にあること、4)「賃貸コンテンツサービス」 (注3) を提供しているプロバイダはWebの世界で中規模なハブサイトとして重要な位置を占めていること、などがある。

一方で、本研究は全体的な特徴を調査することに重点を置いたため、個々のサイトに関しては詳しく解析していない。また提供されている個々の情報内容に関しても解析していない。実際の個々の情報要求に直接的に対応するには、提供されている情報の内容分析をする必要もある。

本研究で重要であると判明した個々のサイトに対して、例えば、1) どのような情報を提供しているのか (Yellow Pageか1次情報か)、2) 主に提供している情報の主題あるいは分野、3) 他のサイトから多く参照されて

いる情報 (セールスポイント) は何か、などの内容分析を行なうと、統計的な判断に基づいた重要な情報 (ページ) のみを見つけることができる。これらのページを体系的にまとめることにより、多くの利用者が欲すると思われる重要なもののみを集めた、Yellow Page を作成することも可能である。本研究はこのように Web の全体的な特徴を解析することにより、Web の情報検索支援のための基礎データを与えることもでき、統計データに基づいた情報検索支援ができる可能性をも秘めている。

注 1: 1996 年 10 月 1 日に JICST と新技術事業団 (JRDC) とが統合され、科学技術振興事業団 (JST) となった。現在はこの団体名鑑のページ群は大幅に規模が縮小されており、残念である。現在の URL は <URL: http://www3.jst-c.go.jp/Inst_dir/>

注 2: URL Square (ORIONS), Available from <URL: http://www.orions.ad.jp/urls/index-jp.html>.

注 3: プロバイダの多くは、ユーザが自分の情報を Web 上で情報公開を行なえるようなサービスを展開している。ここではこのサービスを「賃貸コンテンツ提供サービス」と呼んだ。このサービスを利用して (独自にインターネットに接続せずに)Web 上で情報を公開している組織も多数ある。今後このサービスを利用して情報提供を行なう組織がさらに増加すると、解析の際にこれらのプロバイダの特別扱いが必要となるだろう。

参考文献

[1] 中川格. World Wide Web 情報空間の特徴の分析と把握. 修士論文, 図書館情報大学, 1997. Available from <URL: http://voyager.ulis.ac.jp/papers/thesis/>.

[2] Nakagawa, Itaru et al. An analysis of Internet resources: Toward drawing a WWW server relationship map. Proceedings of Fifth Conference of International Federation of Classification Societies 96, Kobe, 1996-03, Internatinal Federation of Classification Societies. Vol. 1, p77–80(1996). Available from <URL: http://voyager.ulis.ac.jp/papers/abs-IFCS96.ps>.

[3] 日本ネットワークインフォメーションセンター. JP ドメイン名の割り当てについて. 1996-11-06. Available from <URL: ftp://ftp.nic.ad.jp/pub/jpnic/domain-name-all.txt.961106>.

[4] 日本ネットワークインフォメーションセンター. JP ドメイン名 (地域型) 割り当てについて. 1996-08-05. Available from <URL: ftp://ftp.nic.ad.jp/pub/jpnic/domain-geographic.txt>.

[5] 日本ネットワークインフォメーションセンター. 日本ドメイン名一覧表. 1996-06-08. Available from <URL: ftp://ftp.nic.ad.jp/pub/jpnic/domain-list.txt>.

[6] 田村健人. Senrigan search. Available from <URL: http://www.info.waseda.ac.jp/search.html>.

[7] Woodruff, Allison et al. An investigation of documents from the World Wide Web. Proceedings of Fifth International World Wide Web Conference, Paris, 1996-05. Available from <URL: http://www5conf.inria.fr/fich_html/papers/P7/Overview.html>.

表 1: 日本の Web サーバ数とサイト数

Groups	The Numbers of Web Servers [A]	The Numbers of Web Sites [B]	Internet Connected Sites [C]	Web Servers per Web Site [A/B]	Web Sites percentage [100×B/C]
ac.jp	6614 (59.0%)	576 (16.2%)	656	11.5	87.8
ad.jp	184 (1.6%)	77 (2.2%)	95	2.4	81.1
co.jp	2802 (25.0%)	2187 (61.4%)	3725	1.3	58.7
go.jp	493 (4.4%)	153 (4.3%)	180	3.2	85.0
or.jp	655 (5.9%)	338 (9.5%)	1073	2.0	31.5
others	456 (4.1%)	229 (6.4%)	330	2.0	69.4
total	11204 (100%)	3560 (100%)	6059	3.2	58.8

表 2: リンクの向きによる種類分け

Types	Significance
Internal Links (Int. Links)	同一サイト内の http オブジェクトを参照しているリンク
External Outgoing Links (Ext. Out. Links)	他サイトの http オブジェクトを参照しているリンク
External Incoming Links (Ext. Inc. Links)	他サイトの http オブジェクトからのリンク

表 3: Active Sites の数とその割合

Groups	Active Sites [A]	Web Sites [B]	Active Sites per Web Sites [100×A/B]	Percentage in Active Sites [100×A/C]
ac.jp	300	576	52.1	24.3
ad.jp	45	77	58.4	3.7
co.jp	520	2187	23.8	42.1
go.jp	62	153	40.5	5.0
or.jp	210	338	62.1	17.0
others	98	229	42.8	7.9
Total	[C]... 1235	3560	34.7	100

表 4: 3 種類のリンクの量

Groups	Int. Links [A]	Ext. Out. Links [B]	Ext. Inc. Links [C]	[100× B/A]	[100× C/A]
ac.jp	2,382,222	350,144	223,367	14.7	9.4
ad.jp	213,089	202,601	98,569	95.1	46.3
co.jp	1,418,443	40,822	159,559	2.9	11.2
go.jp	591,330	17,266	28,017	2.9	4.7
or.jp	454,074	46,268	155,709	10.2	34.3
others	89,575	39,583	31,463	44.2	34.1
Total	5,148,733	696,684	696,684	13.5	13.5

表 5: グループ間クロスリファレンス数 (式 4)

from Groups	to Groups						
	ac.jp	ad.jp	co.jp	go.jp	or.jp	others	Total
ac.jp	151,246	30,915	72,234	15,152	60,348	20,249	350,144
ad.jp	38,813	61,217	41,551	6,177	49,680	5,163	202,601
co.jp	7,729	1,951	16,010	1,214	12,546	1,372	40,822
go.jp	6,325	857	3,931	2,334	2,679	1,140	17,266
or.jp	12,444	2,146	12,599	1,861	14,877	2,341	46,268
others	6,810	1,483	13,234	1,279	15,579	1,198	39,583
Total	223,367	98,569	159,559	28,017	155,709	31,463	696,684

表 6: Internal Links の平均 (式 5) とクロスリファレンスの平均 (式 6)

from Groups	Int. Links	to Groups						
		ac.jp	ad.jp	co.jp	go.jp	or.jp	others	Total
ac.jp	7940.7	504.2	103.1	240.8	50.5	201.2	67.5	1167.1
ad.jp	4735.3	862.5	1360.4	923.4	137.3	1104.0	114.7	4502.2
co.jp	2727.8	14.9	3.8	30.8	2.3	24.1	2.6	78.5
go.jp	9537.6	102.0	13.8	63.4	37.6	43.2	18.4	278.5
or.jp	2162.3	59.3	10.2	60.0	8.9	70.8	11.1	220.3
others	914.0	69.5	15.1	135.0	13.1	159.0	12.2	403.9

表 7: 3 種類のリンクの量の中央値

Groups	Int. Links	Ext. Out. Links	Ext. Inc. Links
ac.jp	1247	132	211
ad.jp	376	40	142
co.jp	434	4	96
go.jp	504	17	113
or.jp	530	23	101
others	112	5	49
total	499	14	108

表 8: グループ間平均参照率 (式 9)

from Groups	to Groups					
	ac.jp	ad.jp	co.jp	go.jp	or.jp	others
ac.jp	16.4	11.2	8.3	11.4	7.7	5.7
ad.jp	6.2	7.1	3.2	5.8	4.1	3.2
co.jp	1.4	1.8	1.5	1.6	1.6	0.7
go.jp	8.2	7.6	4.7	13.7	4.3	5.1
or.jp	3.3	3.7	3.0	3.4	3.7	2.4
others	3.0	2.6	1.4	2.8	1.8	4.3

表 9: グループ間相互参照率 (式 14)

from Groups	to Groups					
	ac.jp	ad.jp	co.jp	go.jp	or.jp	others
ac.jp	30.5	14.8	6.4	18.5	12.5	16.6
ad.jp	–	10.0	5.9	10.7	8.1	16.4
co.jp	–	–	6.6	6.0	7.1	10.8
go.jp	–	–	–	30.0	10.1	12.3
or.jp	–	–	–	–	10.2	17.5
others	–	–	–	–	–	9.3

表 10: Best 50 サイト

Ranking	サイト名	組織名	スコア
1	u-tokyo.ac.jp	東京大学	80.2
2	ntt.jp	日本電信電話株式会社	78.1
3	keio.ac.jp	慶應義塾大学	76.0
4	kyoto-u.ac.jp	京都大学	73.0
5	tut.ac.jp	豊橋技術科学大学	72.7
6	osaka-u.ac.jp	大阪大学	72.3
7	chiba-u.ac.jp	千葉大学	71.5
8	affrc.go.jp	農林水産省農林水産技術会議	69.8
9	sut.ac.jp	東京理科大学	69.7
10	nagoya-u.ac.jp	名古屋大学	69.7
11	jicst.go.jp	日本科学技術情報センター	69.2
12	titech.ac.jp	東京工業大学	68.6
13	orions.ad.jp	大阪地域大学間ネットワーク	68.5
14	tohoku.ac.jp	東北大学	68.2
15	asahi-net.or.jp	朝日ネット	68.1
16	hokudai.ac.jp	北海道大学	67.8
17	uec.ac.jp	電気通信大学	67.4
18	kyutech.ac.jp	九州工業大学	66.7
19	tsukuba.ac.jp	筑波大学	66.7
20	nikkeibp.co.jp	日経 BP 社	66.5
21	aist-nara.ac.jp	奈良先端科学技術大学院大学	66.4
22	saitama-u.ac.jp	埼玉大学	66.3
23	tokai-ic.or.jp	東海インターネットワーク協議会	65.7
24	etl.go.jp	電子技術総合研究所	65.4
25	sony.co.jp	ソニー株式会社	65.2
26	infoweb.or.jp	InfoWeb	65.2
27	tokushima-u.ac.jp	徳島大学	65.1
28	kobe-u.ac.jp	神戸大学	64.8
29	iijnet.or.jp	IIJ インターネット	64.5
30	waseda.ac.jp	早稲田大学	64.1
31	jaist.ac.jp	北陸先端科学技術大学院大学	63.3
32	kyushu-u.ac.jp	九州大学	63.0
33	bekkoame.or.jp	株式会社ベッコアメ・インターネット	62.7
34	nitech.ac.jp	名古屋工業大学	62.4
35	okayama-u.ac.jp	岡山大学	62.2
36	rim.or.jp	リムネット	61.6
37	hiroshima-u.ac.jp	広島大学	61.5
38	mainichi.co.jp	株式会社毎日新聞社	61.0
39	gunma-u.ac.jp	群馬大学	60.4
40	sphere.ad.jp	InfoSphere	60.3
41	ulis.ac.jp	図書館情報大学	59.8
42	ynu.ac.jp	横浜国立大学	59.5
43	yamanashi.ac.jp	山梨大学	59.5
44	ncc.go.jp	国立がんセンター	59.4
45	shinshu-u.ac.jp	信州大学	59.2
46	yamaguchi-u.ac.jp	山口大学	59.0
47	toyama-u.ac.jp	富山大学	58.7
48	anchor-net.co.jp	アンカーネットワークサービス	58.6
49	kagoshima-u.ac.jp	鹿児島大学	58.6
50	impress.co.jp	株式会社 インプレス	58.0

表 11: Best 50 サイトのアクティビティ

リンク数		平均参照率	相互参照率
Best 50 間リンク数	全体に対するシェア率		
2,556,355	43.7%	91.5%	83.6%

表 12: 第 2 レベルドメイン名に基づいたグループわけ

グループ	組織形態	該当組織	使用されている登録ドメイン数
ac.jp	教育および学術機関 [大学]	学校教育法および他の法律の規定による学校 (小・中学校、および高等学校を除く)、学校法人、大学共同利用機関、大学校、職業訓練法人	656
ad.jp	JPNIC 会員 [ネットワーク管理組織]	JPNIC 会員ネットワーク、JPNIC がネットワーク運用上必要と認めた組織	95
co.jp	企業 (または営利法人)	株式会社、有限会社、合名会社、合資会社、相互会社、特殊会社、その他の会社および信用金庫、信用組合その他の営利法人	3725
go.jp	日本国政府機関	政府機関、各省庁所轄研究所、特殊法人 (特殊会社を除く)	180
or.jp	団体	財団法人、社団法人、宗教法人、監査法人、その他 ac.jp,co.jp,go.jp に属さない法人、任意団体、外国政府機関の在日公館その他の組織ならびに、国連,EU 等の国際的公的機関、各国地方政府 (州政府) の駐日代表部事務所	1073
others	その他	地域型ドメインと、"ntt.jp" などの特殊なドメイン	330

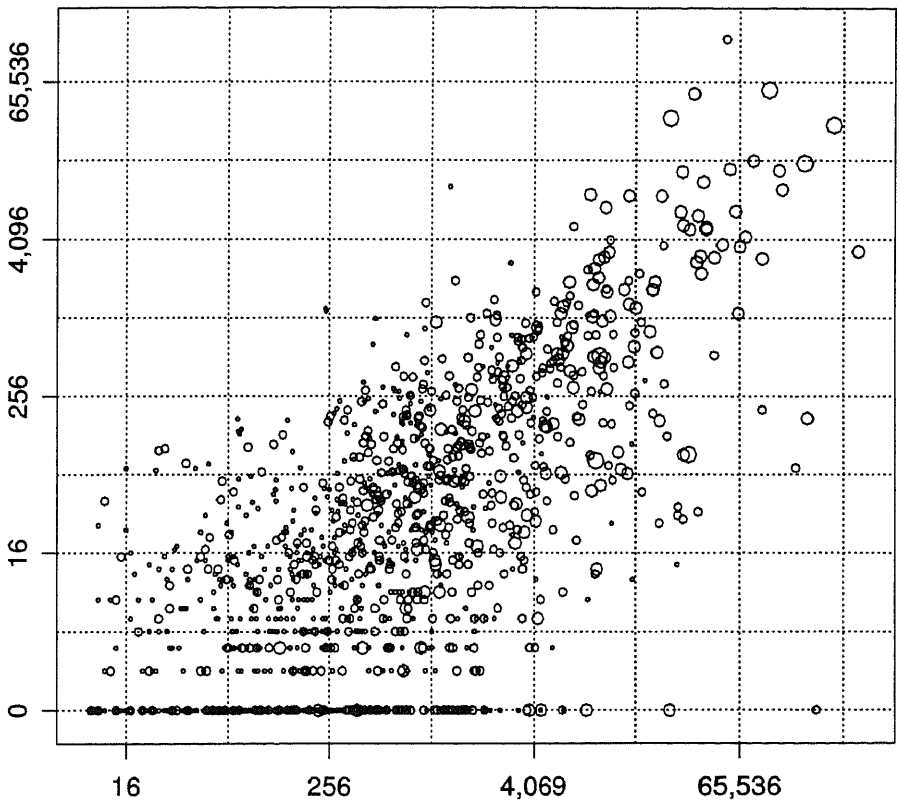

図 1: Int. Links[X] と Ext. Out. Links[Y] に対する Ext. Inc. Links の量 (円の直径)

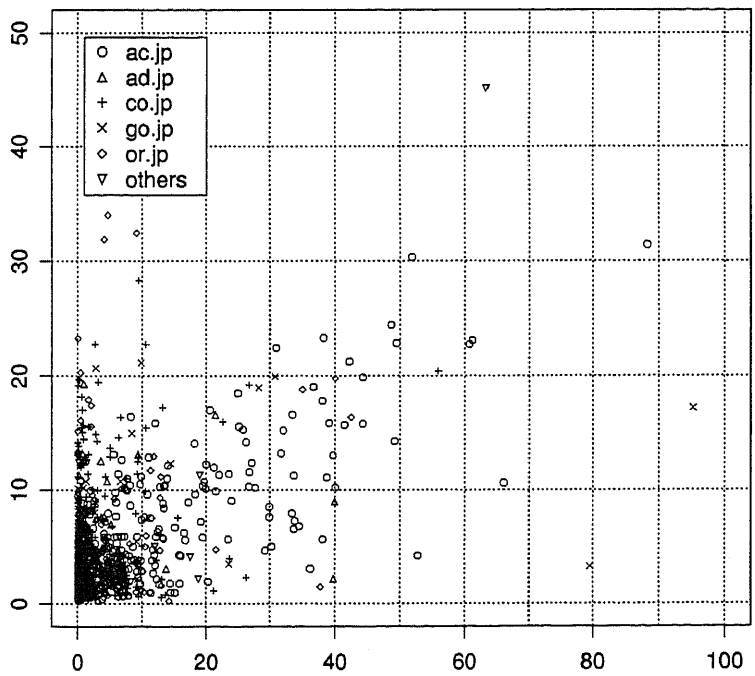

図 2: 参照率 (X) と被参照率 (Y)

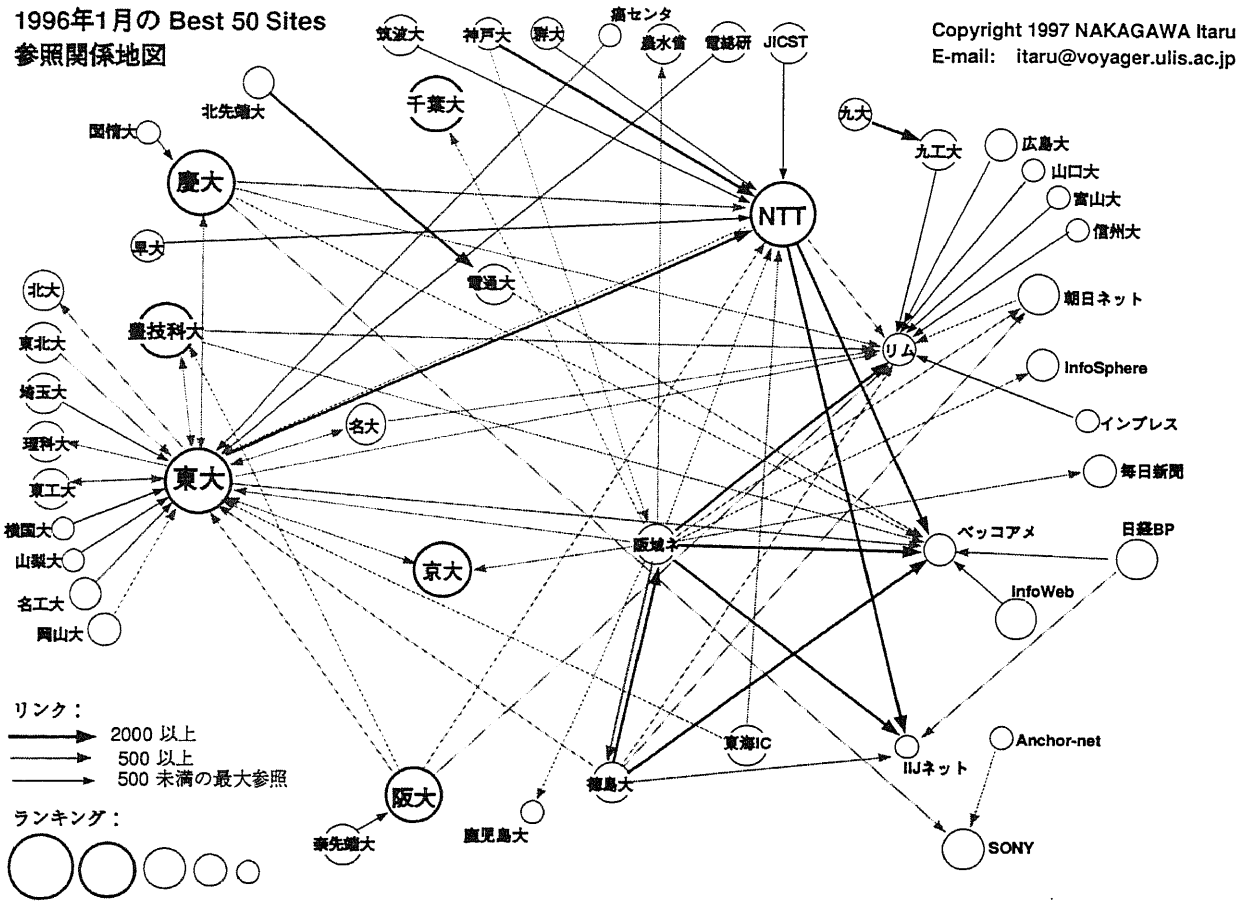

図 3: Best 50 サイト参照関係地図

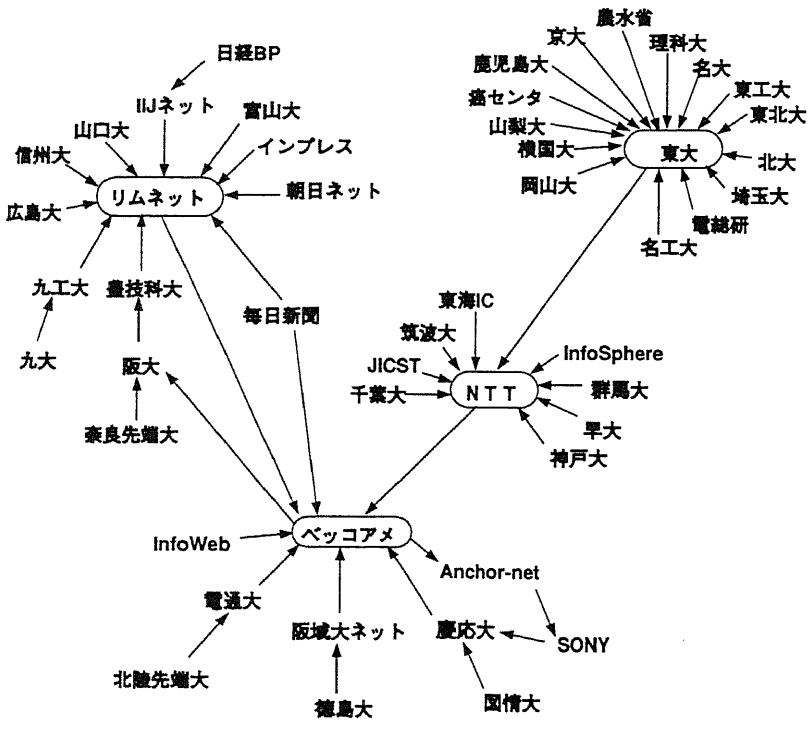

図 4: 最大参照先による地図