ヒューマンインタフェース 70－3
（１９９７．１．１６）

ユーザの利用履歴に基づくWWWサーバの地図型ディレクトリ

早川 和宏　　福永 博信　　鈴木 達郎
{hayakawa, fuku, tsuzuki}@aether.hil.ntt.co.jp
NTT ヒューマンインタフェース研究所
〒238-03 横須賀市光の丘１－１

概要

本報告では、WWWサーバの集合が持つ情報相互の関連を、それを参照する複数のユーザからのアクセスパターンを通じて可視化することによって、「WWWサーバの構成する情報空間の地図」を自動生成する手法を提案する。本手法はproxyサーバのログファイルからWWWサーバの特徴ベクトルを生成し、主成分分析を行なって多次元空間内でのWWWサーバの布置を3次元で可視化する。本手法で生成された地図は、ユーザから見たWWWの状態を反映する、ユーザ適応型のディレクトリとして利用することができる。

A global map of the WWW auto-generated from users' access history

Kazuhiro Hayakawa　　Hironobu Fukunaga　　Tatsuo Suzuki
{hayakawa, fuku, tsuzuki}@aether.hil.ntt.co.jp
NTT Human Interface Laboratories
1-1 Hikari-no-oka, Yokosuka, Kanagawa 238-03 Japan

Abstract

In this paper, a method for auto-generating a map of the WWW servers is proposed which visualize the relationships between WWW servers by observing the access patterns from users. Our method generates a feacture vector for each WWW server by analyzing an access log file of a proxy server. The layout of WWW servers in multidimentional space is visualized as a three dimentional world using principal component analysis. The layout reflects the users' view of the relationships between WWW servers so that it could be used as an adoptive WWW directory.

1 はじめに

本報告では、WWW が作り出している、日々変化する情報空間の「利用者から見た姿」を可視化し、情報空間の地図として用いる手法について述べる。

本研究は WWW が作っている情報の空間が、利用者にはどのように認識されているかを可視化し、利用者にフィードバックすることで、利用者の「WWW 空間の認知地図」の構成を助け、同種のサーバの探索や、周囲では頻繁にアクセスされているのに普段あまりアクセスしていないサーバの発見、といった効果を生み出すことを目標としている。

2 UI と情報可視化

2.1 情報可視化

情報可視化 (Information Visualization) とは、情報という、普通の三次元空間内では幾何学的実体を持たないものを、画面というユークリッド空間上にマッピングすることである [1][2]。

科学における可視化 (Scientific Visualization) の可視化の対象 (たとえば流れ場など) は通常物理的実体を持っており、それゆえ通常は幾何学的マッピングは自明であることが多い。そのため、可視化の技術的興味の焦点は幾何学的配置を生成することそれ自体よりも、むしろ物理現象を把握しやすくする点に重点がある。

一方、情報可視化においてもデータ構造やデータの変化を把握しやすくすることは大きな目的であるが、もともと現実の空間に存在していない、数学的な実体を可視化するため、それ以前の段階として幾何学的配置の生成手法が問題となる。

多くの研究では、情報の内容をユーザがどう捉えているかは直接着目せず、データ構造の幾何学的配置を、あらかじめ決めた評価関数に従って最適化するものが多い。例えばグラフの可視化の研究 [3] では、ノードの配置を決定するために、配置から計算されるポテンシャルをもっとも大きくするような配置を効率的に求めるといったことが行なわれる。

これに対し著者らは、情報の可視化において重視すべきなのは、情報が計算機上でどのような構造になっているかよりも、情報がユーザにどのように認識されているかであり、可視化された結果とユーザの認知モデルとが対応することが重要であると考えている。

例えば感性工学の研究やマーケティングにおいては、商品や企業の「消費者イメージ」を調査することがあるが、情報可視化においても情報の消費者イメージを可視化することにより、情報を直観的に把握できるような形で可視化することができるのではないだろうか。

ユーザインタフェースにとっての情報可視化の適用分野としては、情報可視化はユーザの認知的負荷を軽減するという観点から、インタフェースデザインの問題と関連して扱われることが多い。この方向での情報可視化の主要なアプリケーションとしては、情報検索インタフェースや情報ブラウジングのインタフェースがある [4]。たとえばツリー構造を可視化する ConeTree[5] などの研究が有名である。単純な可視化ではなく、利用者に適応した可視化を行なうような拡張 [6] も行なわれている。

また、可視化された情報自体を新しい情報としてより積極的に用いる例としては、思考や議論の発展状況を可視化して、発想支援やグループワークに用いるという研究 [7] もある。

2.2 Hypretext における情報可視化

WWW は Hypertext システムの一種であるので、WWW の可視化も Hypertext の可視化の一種である。

Hypertext においては、Hypertext の全体的な構造を可視化して利用者に見せる、いわゆる「グローバルマップ」が迷子問題の解決などに有用であるとされてきた。しかし有効なグローバルマップを生成するための決定打と言える方法はまだ存在していない。

グローバルマップの生成法として最も直観的な方法は、グローバルマップ上でのノード間の距離が、ノード同士のリンクのされ方を反映するような方法であろう。

たとえば [8] で述べられている minimal links placement strategy は、二つのノード間を結ぶ経路のうち最もリンクの数が少ないもののリン

クの数をノード間の距離とする。

この定義は、最短経路がすなわち（日常的な意味での）距離であるとするという、直観的にはわかりやすい定義である。しかし、この定義では、（数学的な意味での）距離における三角不等式[1]が成り立たないため、これをそのままユークリッド空間の中での距離として考えることはできない。

この定義の元で測定された距離をできるだけ反映するように、グローバルマップを作成することは可能である。しかし、この距離の定義は新たなリンクの追加に対して不安定である。例えば、この定義では新しいリンクを一つ追加することにより、それまでのノード同士の距離が大きく変化してしまう。

従って、常に変化するWWWのようなHypertextで、地図という幾何学的な出力を得るためには、この手法は安定な地図を得ると言う点では都合が悪い。

また、リンクは個々のページの作成者がその作者の観点から作成したもので、利用者の観点から作成されたものではない。従って、リンクにはそれぞれ利用者にとってどのくらい重要かという重みがあるはずであるが、これもHypertextのリンク情報からは得られない。

本報告では、ノード毎に特徴ベクトルを定義し、ノード間の距離は特徴ベクトル同士の距離を用いている。特徴ベクトルを用いる手法は三角不等式を満たしており、上に述べたような不安定さはない。

3 WWWサーバ空間の可視化

3.1 特徴ベクトルの生成

著者らはTCP/IPのproxyサーバであるsockdのログファイルからWWWサーバの特徴ベクトルを得る方法についてすでに[9]で報告している。本報告で用いるログファイルは[9]と同様に1994年5月～7月の間に著者らの属するNTTの研究所で記録されたもので、64人のユーザと335個のサーバが入っている。

[1]三点A, B, Cがあるとき、$AB \leq AC + BC$であること。

この方法では、ログファイルから「複数のユーザ」×「複数のサーバ」の特徴マトリクスが得られる。すなわち、m個のサーバとn個のユーザ（クライアント）が登場するログファイルから、$m \times n$の特徴マトリクスが得られ、マトリクスの(i,j)の要素には、i番めのサーバとj番めのユーザの関連を表す特徴量が入る。

本報告では、特徴量の値をクライアントjからサーバiへのアクセス回数とした。ただし、単純にTCPでの接続の回数をアクセス回数として用いると、ページ中のインライン画像の取得などもすべて1回のアクセスと見なされてしまう。これらのアクセスは全てクライアントソフトが自動的に行なうもので、ユーザの行動とは無関係である。

そこで、今回「同一ホストへの連続したアクセスは、30分以内の間隔ならば何回あっても1回のアクセスと数える」という規則（以下「30分基準」と呼ぶ）のもとでのアクセス回数を測定した。

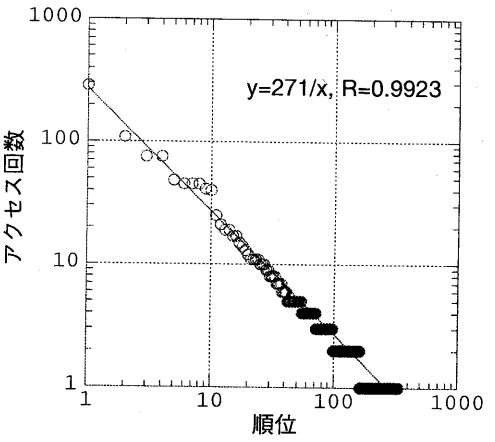

図1: アクセス回数の分布

図1は、30分基準の下で計測した、各サーバの被参照回数を、回数の多い順に並べたものである。x, y軸とも対数になっているが、ほぼ

$$y = \frac{271}{x}$$

の分布に従っている。$y = C/x^p$はZipfの分布と呼ばれる分布で、もともと単語の文書中での出現頻度の分布であるが、他にもレンタルビデ

オの貸し出し頻度など、さまざまな局面で観測される分布である。

一方、WWW とは別の Hypertext システムにおいて、ノードの被参照回数が Zipf の分布に従うことが示されている [10]。今回の 30 分基準においても同様の分布が得られていることは、30 分基準が実質的なアクセス回数を計測する上で有効に働くことを示唆している。

以上のようにして、proxy サーバのログファイルからサーバをとユーザの関連を表す特徴マトリクスを作成した。

3.2 特徴ベクトルの主成分

ログファイルから得られた特徴ベクトルを主成分分析することにより、アクセスパターンに類似性があるユーザ同士を主成分にまとめることができる。同時に、そのような主成分で表される、あるコミュニティからアクセスされているサーバ同士もまとめることができる。

主成分分析の結果、多数のユーザからなる多次元空間内でのサーバの位置を、少ない主成分の結合で近似的に表すことができたとする。すると、主成分で表される複数のコミュニティのそれぞれからどれだけアクセスされているかを基準にして、サーバをより少ない次元の空間の中に再配置することができる。

前節で得られた特徴マトリクスに主成分分析を行なった結果を次の表に示す。元のデータに平均値を引いたり標準偏差で割るという操作は加えていない。

主成分	固有値	%	累積%
1	16.803	38.4	38.4
2	10.959	25.0	63.4
3	5.9506	13.6	77.0
4	2.1118	4.8	81.9
5	1.5441	3.5	85.4
6	1.3924	3.2	88.6
7	0.77435	1.8	90.3
8	0.66512	1.5	91.9
9	0.51541	1.2	93.0
10	0.49906	1.1	94.2

表から分かるように、第三主成分までで全分散の 77% を説明している。ただし、常に三つの主成分でこの程度の割合の分散を説明できるとは限らなず、第三主成分までで 77% という値はログファイル中に出現するユーザの総数（今回は 64 人）に依存している可能性もあるが、今回のデータに限っては、第三主成分までに注目しても元の空間の状態をある程度表現することができる。

3.3 サーバ空間の可視化

特徴ベクトルを主成分分析して得られた主成分から、2 つあるいは 3 つの主成分を取り出し、これらを座標軸とする空間内に各サーバを配置することにより、サーバの地図を作成することができる。空間内でのサーバの布置は多次元空間内での布置を二次元や三次元に射影したものとなる。

従って、利用者はあるサーバと似通ったユーザを持つ別のサーバが近くに配置されているような地図を手にすることになるはずである。

しかし、三つの主成分をそのまま軸に取って点の分布を見ると、アクセス数が少ない多くのサーバが原点付近に集中する、見づらいものとなってしまう。これはアクセスされる回数がサーバによって大きく異なっているためである。

ここで、得られた布置について考えてみると、元の特徴量にアクセス回数を用いているので、主成分もアクセス回数の分散を説明するような量となっているはずである。従って、原点からの距離はアクセス回数の総計に関係する。

一方、ユーザ集団の特徴は主成分の違いとして表され、その特徴は点が原点から見てどの方向にあるかに反映される。

従って、サーバの特徴をよく表すのは原点からの距離ではなく、原点からみて点がどの方向にあるかであると考えられる。

そこで、原点からの距離について正規化を行い、各点が単一の球面上に位置するように点の座標を再決定する。

得られた主成分の第一主成分から第三主成分までを X, Y, Z の値として、各サーバを三次元空間内の点としてマッピングする。今回のデータでは第三主成分までで全分散の 77% を説明しているので、この三次元空間は、元の多次元空間内での点の布置を比較的よく記述すものとなっているはずである。

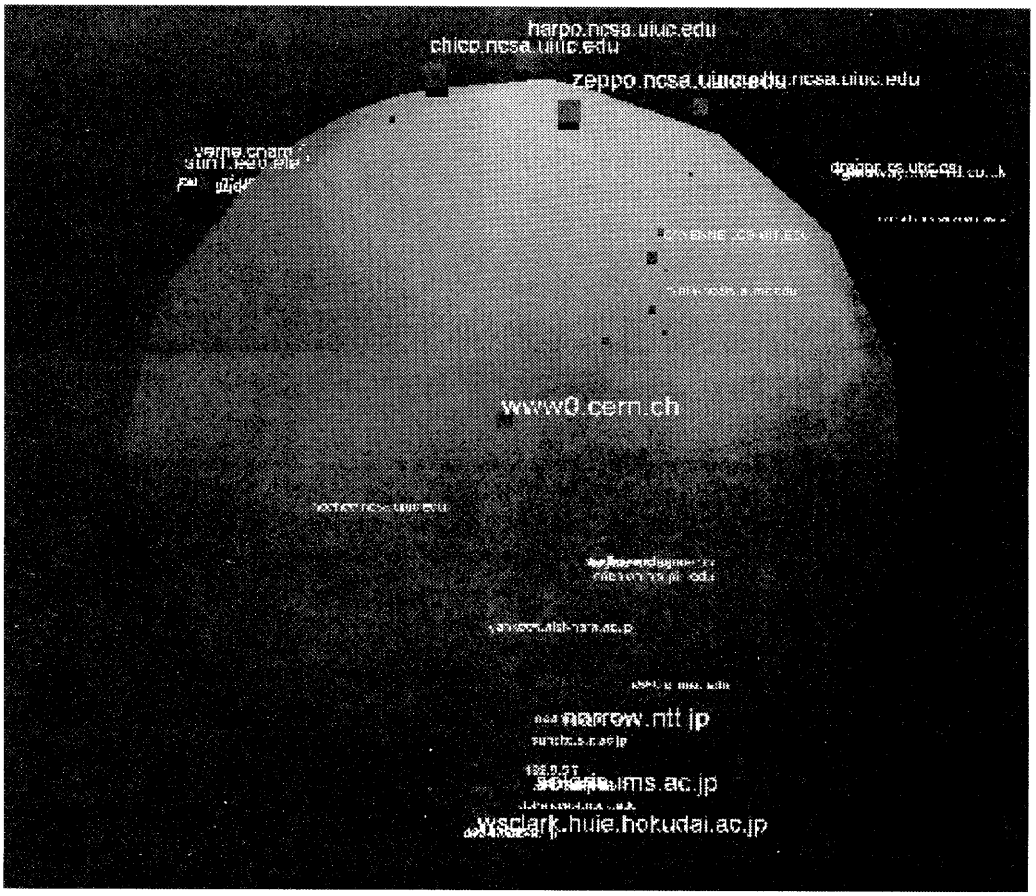

図 2: 可視化された WWW 空間の全体像

4 サーバ地図を用いたインタフェース

4.1 全体像

図 2は前節で説明したように、サーバの特徴ベクトルを主成分分析し、球面上に再配置したデータを VRML[12] で記述したものである。サーバは HTTP アンカーを埋め込んだ立方体で表されており、利用者はクリックすることでそのサーバへアクセスすることができる。また、VRML ブラウザの機能にしたがって自由にこの空間内を探索することもできる。

図で、第一主成分は図の下方向、第二主成分は図の左上方向、第三主成分は図の右上方向となっている。

以下この布置の各部について若干の考察を加える。ただし、ここで挙げた布置はログファイルに現れるユーザの集団（＝著者らの属する NTT の研究所）に依存するものであって、一般的にこのような布置が成立するのではない。全く異なるユーザ集団ではその集団独特の全く異なる布置が得られるはずである。また、ログファイルの時期が古いので、図中に出てくるサーバには現存しないものも多く含まれている。

4.2 第一主成分方向

図 3は、図 2の下の部分（第一主成分方向）を拡大したものである。

この方向には日本のサーバが多く集まってい

る。これはユーザが日本のサーバに限って高い頻度でアクセスしたためと思われる。

図中央の narrow.ntt.jp は当時 NTT ホームページを提供していたマシンで、「What's new in Japan」や「WWW servers in Japan」といったページがあり、日本のサーバのディレクトリとしての役割も果たしていた。

下端近くにある wsclark.huie.hokudai.ac.jp は、Archie Gateway サービスなどを行なっていた北海道大学のサーバである。また、solaris.ims.ac.jp は分子科学研究所のサーバで、ソフトウエアのアーカイブとしても機能していた。

これらのサーバが集まっていることから、第一主成分方向の領域は「日本人＋NTT 社員＋計算機系研究者」といったコミュニティの特徴を反映していると考えられる。

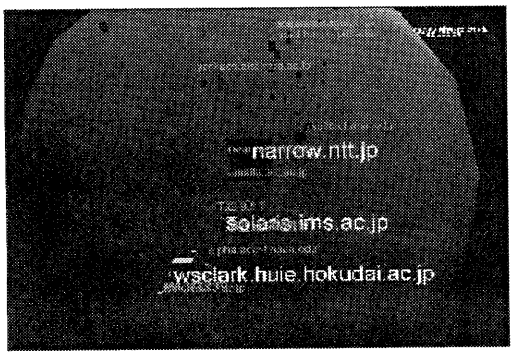

図 3: 第一主成分方向

4.3 中央部分

図 4は、図 2の中央部分（原点から見て (1,1,1) の方向）を拡大したものである。

図の中央にある www0.cern.ch は、WWW の発祥の地として参照されていた CERN のサーバである。ログファイルが記録された当時はまだ WWW が広まりつつある時代で、「WWW とは何か？」というリンクが多くのページにあった。サーバ中のどのページがどのようなリンクから辿られたかの分析は行なえないので、推測であるが、「WWW とは何か？」のリンクがたどられた結果として、このサーバがこの空間の

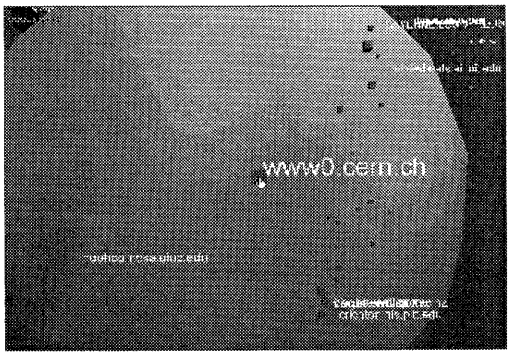

図 4: 中央部分

中央に位置することになったと考えられる。また、CERN のサーバには世界中の WWW サーバをジャンル別に分類した WWW Virtual Library のホームページがある。

(1,1,1) の方向にあるサーバは、三つの主成分の値が比較的均衡していることを示す。これは比較的多くのユーザから偏りなくアクセスされたことを示すと考えられる。従って、中央部分には誰からも均等にアクセスされるようなサーバ、例えばあまりジャンルに偏りのないディレクトリサービスのページなどが集まると考えられる。

4.4 第二・第三主成分方向

図 5は図 2の上の部分を拡大したものである。図中 harpo、zeppo、chico は NCSA Mosaic のデフォルトのホームページである www.ncsa.uiuc.edu の Canonical Name であり、www.ncsa.uiuc.edu がアクセスされると実際にはこれらのホストがアクセスされる。事実上これらのマシンは同一のものと見なせるので、空間中でも近くに位置するのは当然であると言える。

NCSA Mosaic は当時としては日本語の表示できる WWW クライアントソフトとして最もメジャーであったので、むしろもう少し中央部分寄りに位置しても良さそうである。しかし実際には中央から NTT ホームページなどと反対方向に離れた場所へ位置している。これは、第一主成分の「日本人＋NTT 社員＋計算機系研究者」というコミュニティが、早々とホームペー

ジをNTTホームページなどへ切替えたためではないかと思われる。

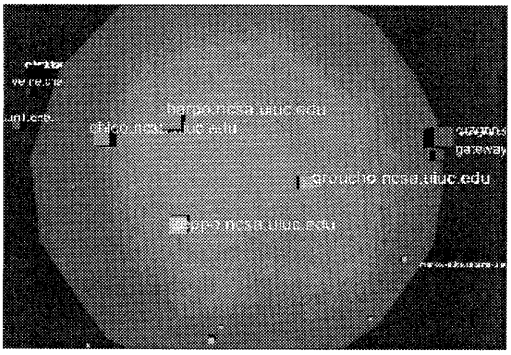

図5: 第二・第三主成分方向

図の左方向は第二主成分の方向で、イメージファイルのアーカイブサイトが数個配置されている。右方向は第三主成分であるが、F1関連情報の置かれていたabekrd.co.ukとカナダのBritish Columbia大学計算機科学部cs.ubc.caのサーバがあるが、意味づけは不明である。

5 考察

本報告では、proxyサーバのログファイルからWWWの情報空間の地図を生成する手法の概略を示した。しかし、細部について検討が不十分なままの点もあるので、それらについて以下に述べる。

一点目は、特徴量としてアクセス回数を用いることの妥当性である。30分基準がWWWへのアクセスの正しいアクセス回数を計るのに有用であることを示したが、アクセス回数をユーザとサーバとの関連を計る尺度として用いることがどの程度正当性があるのかを、今回は検証していない。

特徴量として望ましいのは、ある個人との関連で測定された、各サーバの特徴量を一次元の座標軸上にマッピングするとき、その座標値がある量（たとえばどの程度その個人の嗜好にマッチしているかの度合）と相関しているか、あるいは座標軸上で分類が起こるような量である。

たとえば、ネットワークニュースでは記事の長さに対してどれだけの時間をかけたかが、読者にとってのその記事の重要度と関連するという報告がある[11]。同様にWWWにおいても、サーバがユーザの興味に一致している度合の指標としては、アクセス回数よりはむしろアクセス時間の方が適当であるとする考え方もある。

さらに、同じ一回のアクセスでも、サーバによって情報を取得するのにかかる通信時間（無意味な画像が多いなど）や得られる情報の価値はユーザにとっては同じではない。サーバにとっても、同じサーバへの一回のアクセスでも、何百ものサーバを訪れているユーザからのアクセスと、全部で二つのサーバしかアクセスしていないユーザからのアクセスとを同じ一回のアクセスと見なしてよいかという問題もある。

以上のように、アクセス回数が個人とサーバとの関連を表すと単純に言うことはできない。本報告では主に簡便さの点からアクセス回数を用いているが、本来は利用者にとっての情報の価値を定量的に測定できるようなコンテンツを用意した上で、どのような特徴量が利用者とコンテンツの関連を表すために適当かを調査すべきである。

二点目に、URLの粒度の問題がある。URLはホスト名、ディレクトリ名、ファイル名を含むことができるが、今回は利用したTCP/IPレベルのproxyサーバのログファイルでは、ホスト名しか記録されないため、より細かい粒度での分析が行なえなかった。それでも、このログファイルが記録された当時は、比較的小規模のサーバが多かったために、ある程度サーバを分類することができている。

しかし、現在ではインターネットプロバイダの巨大なサーバに大量の個人ホームページが構築されており、一つのプロバイダでさえ当時のWWWサーバ数よりも多くの個人ホームページを抱えているのではないかと思われる。そのような巨大なサーバは、URLのディレクトリ名以下の部分で区別して、一まとまりの情報として扱う必要がある。

従って、現在のWWWを分析するには、今回のようにTCP/IPのproxyサーバのログではなく、URLを完全な形でログファイルに記録できるHTTPのproxyサーバのログを用いて、より細かい粒度でURLを扱うことが望ましい。また、

proxy サーバのログではなく、巨大な WWW サーバのログファイルを分析すればそのサーバについての地図を作ることも可能と思われる。

6 まとめ

本報告では WWW のログファイルから WWW の地図を作成する手法と、その例を示した。本手法によって作られた地図は、ユーザのアクセス傾向を用いることにより「ユーザの目から見た」WWW の姿を反映したものとなる。

本手法で作成される WWW の地図は、ユーザの集団が持つ性格を反映するものになる。従ってそのユーザ集団に専用の地図が自動的に作られるという点が重要である。

今後の展開としては、一定期間ごとに地図を最新のログファイルから再構成することによって、時間軸に沿ったアクセス傾向の変化を WWW の情報空間の時間的な変化として可視化することが考えられる。

時間軸に沿った変化を見ることにより、WWW の世界の変化やユーザコミュニティの変化のダイナミズムを直観的にとらえることも可能となるだろう。

謝辞

討論して頂いたヒューマンインタフェース研究所映像処理研究部の皆様に感謝致します。

参考文献

[1] Proceedings of the 1995 Workshop on Information Visualization, IEEE Computer Society Press, 1995.

[2] 小池：ビシュアライゼーション, bit 別冊ビジュアルインタフェース, 共立出版, 1995.

[3] 鈴木、鎌田、榎本：単純無向グラフ自動描画アルゴリズム, コンピュータソフトウエア, Vol.12, No.4, 1995.

[4] Card, S. K. : Visualizing Retrieved Information: A Survey, IEEE Computer Graphics and Applications, Vol. 16, No. 2, pp.63-67, 1996.

[5] Robertson, G. G. et al. : Cone Trees: Animated 3D visualization of hierachical information, Proceedings of CHI'91, pp. 189-194, 1991.

[6] 寺岡、丸山：ユーザの視点に基づく適応型3次元インタフェース, 電子情報通信学会研究報告 MVE96-52, 1996.

[7] 角、西本、間瀬：グループディスカッションにおける話題空間の可視化と発言エージェント, 情報処理学会研究報告・情報学基礎 43-15, 1996.

[8] Utting, K., Yankelovich, N. :Context and Orientation in Hypermedia Networks, ACM Transactions on Information Systems, Vol. 7, No. 1, pp.58-84, 1989.

[9] 早川, 鶴巻, 浜田： WWW の利用履歴に基づく WWW サーバの類似検索, 情報処理学会研究報告・情報メディア 21-2, 1995.

[10] Qiu, L. :Frequency Distributions of Hypertext Path Patterns: A Pragrmatic Approach, Information Processing & Management, Vol. 30, No. 1, pp.131-140, 1994.

[11] Morita, M., Shinoda, Y.:Information filtering based on user behavior analysis and best match text retrieval, Proceedings of SIGIR'94, pp.272-281, 1994.

[12] Pesce, M. :VRML, New Riders Publishing, 1995.

[13] 大隈・ルバール：記述的多変量解析法, 日科技連, 1994.