Web 空間における人間関係を用いた情報探索の一手法

風間一洋[†,††] 佐藤進也[†] 福田健介[†]
村上健一郎[†††] 川上浩司[††] 片井修[††]

本論文では，実世界の実体と Web 空間の固有表現を結びつけることで，利用者の検索過程を支援する実世界指向の情報探索手法を提案し，固有表現として人名を用いた例を示す．本手法では，まずあるトピックに関連する Web ページをサーチエンジンで検索し，その検索結果の上位 n 件から人名を抽出して，各 Web ページに対して，ページ内の出現位置を考慮して共起関係を解析する．次に，新たな中心性の尺度である影響度に基づいて選択することで，あるトピックに関する中心的な人物から構成される人間関係を求めて，Web 空間の情報探索に使用できるようにする．さらに，実際に収集した Web ページ群に対して，人名の出現分布，各パラメータの変化による人間関係の変化，抽出精度を分析する．最後に，さまざまな検索語に対して，それから求められる人間関係を情報探索に用いる場合の妥当性について考察する．

A Method of Information Navigation by Human Relationships on the Web

Kazuhiro Kazama,[†,††] Shin-ya Sato,[†] Kensuke Fukuda,[†]
Ken-ichiro Murakami,[†††] Hiroshi Kawakami[††]
and Osamu Katai[††]

In this paper, we propose a method to assist user's search process by a real-world oriented searching which combines the named entities on the web and objects in the real world. We show an example using personal names as named entities. For this purpose, we extract personal names from the top n search results which are relevant to a specified topic and analyze their co-occurence on ground of their positions in each web page. We extract the relationships of key people which are relevant to a specified topic using the co-occurence and a new centricity measure called "effectiveness." We use the relationships for information navigation on the web. Moreover, we analyze the distribution of personal names, the human relationships with varying parameters, and the precision of the human relationships. We discuss about the adequacy of human relationships as information navigation paths for various search queries.

1. はじめに

Web の急速な普及にともない，膨大で多種多様な文書が利用できるようになり，サーチエンジンを用いて目的の情報を検索することは，業務や生活に必要不可欠な行為となっている．半面，その情報量の膨大さゆえに**情報の過負荷**（information overload）が発生し，目的とする情報が存在するにもかかわらず，利用者がうまく探し出せないことも多い．これを**ハイパー空間の迷子問題**と呼ぶ．

この原因は，Web 空間の探索手段が，ハイパーリンクによる巡回とサーチエンジンによる検索に限られるからである．前者では，つねに Web ページの著者が指定した順路を通ることになり，情報の網羅性や概観性に欠ける．一方，後者では，情報の網羅性は高くなるが，概観性に欠けるので，膨大な検索結果から目的の情報を見つけ出すためには，目的のトピックに対して適切に検索結果を絞り込むことが重要になる．このための一般的な方法は，複数の検索語で AND 検索することである．実際のサーチエンジンの利用履歴分析によると，日本の ODIN では，平均検索語数は 1.42 語で，1〜2 語が全体の 91.4%を占めており[1)]，海外の AltaVista では，平均語数は 2.35 語で，1〜3 語が全体

† NTT 未来ねっと研究所
NTT Network Innovation Laboratories
†† 京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University
††† 法政大学ビジネススクールイノベーション・マネージメント研究科
Hosei Business School of Innovation Management

の 72.4%を占めていた[2)☆]．また onestat.com の 2004 年 7 月の調査では，より多くの語を用いる傾向がみられたと報告されているが，1〜3 語が全体の 73.52%を占めており，大きくは変化していない[3)]．つまり，利用者の大部分は，検索結果数をうまく絞り込めずに，膨大な検索結果の上位の少しだけしか見ていないと考えられる．高度な検索スキルを持たなくても，うまく情報を探索できるようにするためには，検索結果の概観性を向上させるような支援が必要になる．

本論文では，新たな情報探索手段として，実世界の実体と Web 空間の固有表現を結びつけることで，利用者の検索過程を支援する**実世界指向情報探索**を提案し，実際にその固有表現として人名を用いて，人間関係を利用した情報探索の例を示す．まず，検索結果の Web ページから，人名という実世界の既知のエンティティを指す固有表現を抽出し，それを検索結果と同時に提示して，Web ページの属性の判断，または Web 空間内を探索する際のランドマークとして使用できるようにすることで，概観の把握や絞り込みを容易にする．次に，ある情報に直接関連する人物だけでなく，その人物と Web ページで共起する人物も同時に表示することで，友人の友人関係（Friends of a Friend）に相当する人間関係を用いた，効率的な情報探索を可能にする．

本論文の構成は，以下のとおりである．2 章では，関連研究について述べる．3 章では，人間関係を用いた情報探索について，4 章では，人名および人名共起関係に基づいて人間関係を抽出する手法と 3 つのパラメータについて，5 章では，人間関係を利用した情報探索手法と可視化手法について述べる．6 章では，対象とした Web 空間の人名分布，3 つのパラメータが処理結果に与える影響，得られた人間関係の妥当性，および情報探索における人間関係利用の妥当性について分析する．最後に 7 章で結論を述べる．

2. 関 連 研 究

Web の情報探索時に検索関連語を提示して検索結果の絞り込みや拡大，視点の切替えなどを支援する研究では，検索関連語を Web ページから求める手法と検索履歴から求める手法に大きく分けられる．前者としては，神林らの Ingrid では，Web ページから重要な単語を抽出しておき，検索結果に含まれる単語を利用者に検索関連語として提示し[4)]，Kawano の Mondou では，Web ページにデータマイニング手法を適用して求めた相関ルールを用いて，与えられた検索語から検索関連語を導出した[5)]．後者としては，原田らの ODIN では，得られた検索結果に対して，不特定多数の利用者が利用した検索語と実際に閲覧した URL の相関を分析して検索関連語を求め，検索絞り込みに利用した[6)]．ただし，これらの研究では，一般的すぎて効果的ではない検索関連語を排除しにくい，また効果的だが特殊すぎる検索関連語が検出されにくいという傾向がある．戸田らは組織名，地名，人名などの固有表現に着目し，検索結果の Web ページから**固有表現抽出**（Named Entity Extraction）で得られた検索関連語を絞り込み検索に使用したときに，分類の明確さ，均一さ，網羅性が向上することを示した[7)]．本論文でも，固有表現としての人名を使用しているが，さらにある人名に関連する人名まで同時に表示できるようにした点，さらにその関係構造を用いた情報探索を実現した点が異なる．

また，Web 空間における社会ネットワーク構造に関する研究もいくつか存在する．Kautz らの Referral Web のコンセプト証明版では，与えられた人名をサーチエンジンで検索して得られた検索結果の Web ページの取得と，そのページに含まれている人名の抽出を繰り返して，漸次的・部分的に社会ネットワーク構造を得た[8)]．関係の強さの判定には，Jaccard 係数を用いている．松尾らも，既知の人名リストに対して，それに含まれる各人名をサーチエンジンで検索した後に，その検索結果の上位の Web ページを取得して人名を抽出し，さらにそれらとの人名の組の共起関係を検索することを繰り返すことで漸次的に人間関係ネットワークを抽出した[9)]．関係の強さの判定には閾値つき Simpson 係数を用いており，さらにあらかじめ用意した特徴語が名前が共起したページの中に存在するかどうかで，エッジレベルを抽出している．Ogata らの SocialPathFinder では，与えられた URL の Web ページから，Web ロボットを用いてハイパーリンクをたどって指定されたホップ数まで Web ページを収集してメールアドレスを抽出し，人間関係ネットワークを抽出した[10)]．これに対して我々の NEXAS//KeyPerson は，与えられた検索語で検索した結果の上位の Web ページ集合に対して，各ページに出現する人名集合を求めた後で，その共起関係を解析して社会ネットワーク構造を求めた[11)]．さらに，本論文では，影響度と Web ページ上の人名間の距離を用いた人名選択と人間関係可視化，Web ページ，Web サーバ，人名の相関関係の対話的な把握，ある人名に関連する人名の同

☆ AltaVista の方が平均語数が多いのは，たとえば “search engine” は英語では 2 語でも，日本語では「サーチエンジン」と 1 語の複合語で表されるような言語の違いによる．

時表示，人名のクリックによる人間関係の巡回，さらに人間関係可視化部と検索 GUI 部の密接な連携などの拡張を行った．本手法が他の手法と大きく異なるのは，まず任意のトピックに対して妥当な人名を自動抽出できる点にある．ノイズが多い膨大な Web 情報からの妥当な人間関係の抽出は一般に困難であり，誤った関係が抽出されやすいために，結局 Kautz らは既存の文献データベースを用いる方針に切り替え[12]，松尾らは人工知能学会の研究者のようなあらかじめ用意した人名リストに限定しており，Ogata らはリンクされたごく狭い範囲の Web ページしか扱わない．本手法では，リンク解析に基づくオーソリティ度，サーバという局所性に着目した中心性，Web ページ内の距離などのさまざまな要素を考慮することで，妥当な自動抽出を実現している．また，人間関係を漸次的・局所的ではなく全体的に一括で抽出することから，人間のコミュニティが複数独立して存在する場合も求めることができる．次に，生成される Web ページの数は作成者や組織によって大きく異なり，それに人名の共起数も大きく左右されるが，単なる Jaccard 係数や閾値付き Simpson 係数では，たとえば特定の Web サイトに特定の人名群が多量に生成されているような場合に強く影響してしまう問題がある．本手法では，このような問題を回避するために影響度を導入している．さらに，既存の研究では，人間関係を抽出するのが第 1 の目的であり，いったん抽出した人間関係から Web 空間の情報を再び見るようなことは考えていないが，本研究は人間関係と Web 空間との対応関係を保持し，ある人間に関連する Web 情報を自由に閲覧できる機能を提供する．

3. 人間関係を用いた情報探索

3.1 Web 空間と人間の活動

本論文では，実世界指向の情報探索の 1 つの例として，特に Web 空間における人間の活動に注目する．昨今の Web の急速な普及から，人間の生活と Web 空間が不可分になってきている．人間の活動の結果として Web 上の情報が生み出されることから，Web 空間の情報を解析すれば人間の生活を観測できることになる．

たとえば，注目している分野の人たちが，どのような発言をして，どのような興味を持っているのか，さらにどのような人たちと一緒に仕事をしているのかが判明すれば，その分野をこれから勉強しようとしている場合や，その分野の現状や動向を調査している人が情報を探すために非常に役立つ．また，ある分野に関係している人の名前や人数は，その分野が一般的または特殊なのか，また注目されているかどうかなどの判断の基準に利用できる．特に，著名人や専門家に関しては，彼らの行動が信頼できるだけでなく，その影響力から，その分野が将来どうなるかについても推測できる．

自分あるいは他人の行動を知るために，サーチエンジンで人名や別名を検索する行為は，**エゴサーチ**（ego search）と呼ばれている[13]．たとえば，MSN のアンケート調査では，インターネットで「自分自身」を検索したことがある人は 39%,「疎遠になった友人」が 36%,「家族」が 29%,「昔の恋人」が 17%である[14]．このような目的の情報検索は，今後サーチエンジンの重要な使用法の 1 つになる可能性があるが，Web 空間内で各人がどのような行動をしているかを正確に知るためには，多数の検索結果と関連する情報を閲覧して推測する必要があり，必ずしも容易ではない．

そこで本論文では，Web 空間の中では現実世界の人物は，その固有表現である人名で参照されることに着目して，人名の共起関係を解析して人間関係を抽出し，それを情報探索にも利用することを試みる．

3.2 実世界指向の情報探索

従来の Web 情報検索では，入力した検索式に適合する Web ページを検索し，利用者はその検索結果の Web ページからリンクでつながれている Web ページを閲覧する．一度に閲覧できるページは少数に限られており，目的の情報を探索するためには，検索結果を繰り返し取得して見るか，検索式を手直しして再検索するかである．

これに対して，本論文では Web ページから自動抽出された固有表現を提示することで，実空間のエンティティ側から Web 空間を見る新しい視点の提供を試みる．**図 1** に，実世界指向の情報探索の概要を示す．

すでに述べたように，本論文では実空間のエンティティとして人間に着目し，固有表現として人名を使用する．複雑な検索式を使って検索結果を絞り込むのは一般に困難でも，人間は個性のある理解しやすい存在であることから，人間を Web 空間におけるランドマークとして利用できれば，関連する Web ページの内容の傾向を推測したり，情報の取捨選択を効率的に行えると考えられる．なお，厳密にいえば，同姓同名が存在するために人名と人間の関係は必ずしも 1 対 1 とは限らない．本論文では，扱う人名は姓と名の組に限定していること，日本人の名前は同じ発音でも複数の漢字表記が存在すること，そして本手法では検索式によってトピックを限定したうえで，さらにキーパーソンだけを扱うことから，同姓同名の出現確率が低くな

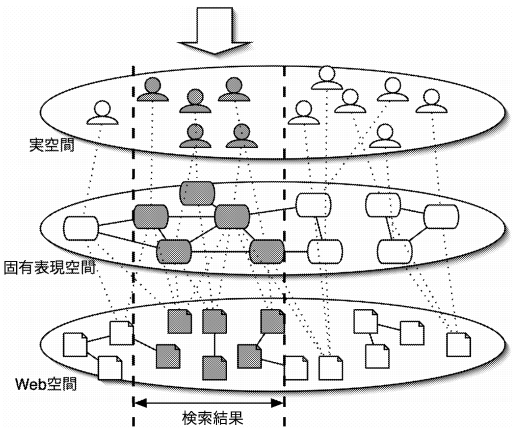

図 1 実世界指向の情報探索手法
Fig. 1 Real-world oriented information navigation.

り，大きな問題にはならない．しかし，英語の Web ページの場合やトピックを特定しない場合には問題が深刻になることが考えられるので，Web 空間における人名の偏在性に着目した同姓同名分離の研究も並行して行っている[15]．

さらに，単に人名を提示するだけでなく，人間関係を求めて，それを新たな情報探索経路としてハイパーリンクと併用できるようにする．たとえば，Milgram や Granovetter は実社会において人間関係が情報探索経路として用いられることを示し[16),17)]，野島らはコンピュータネットワーク上の情報探索の手段として人間関係が用いられることを示している[18)] ように，人間関係は重要な情報探索経路として知られている．さらに，人間関係は “small world” と呼ばれるノード間の平均距離が小さいネットワーク構造を持つことから[19)]，効率的な情報探索経路でもあると推測される．

ただし，Web 空間における人間の活動に関する情報だけを対象とすることから，対象とする情報は限定されることになるが，固有表現は人名，書籍名，地名，店舗名など多岐にわたり，それぞれに対して同じ手法を適用することで，より多種多様な情報を対象とすることができる．逆に，人間の活動に関する情報に限定できるということは，近年さかんに開発されている特定分野に対する検索サービスのように専門的な情報探索が可能になるということであり，それを注目する固有表現を切り替えるだけで同様に使い分けられるとしたら，それは利点であるといえる．

3.3 情報探索が効率化される例

本手法で情報探索が効率化されると思われる例を次にあげる．実際の情報探索は，以下の方法を複数組み合わせて行われる．

- 同じ “XML” の検索結果でも，仕様作成者，アプリケーション作成者などの立場の違いによって人名が出現する情報の内容は異なるので，人名から内容を推測し，情報を取捨選択しやすくなる．
- 後述する影響度が高い，または多くの人間と関係がある人物は重要なキーパーソンであり，特に重要な情報と関係がある可能性が高いので，そのような人物から情報探索を始めると目的の情報が見つかりやすい．
- 人間という既知あるいは理解しやすい存在を Web 空間におけるランドマークとして使用することで，自分が Web 空間のどこにいて，どの部分を調べているかが分かりやすくなり，効率的な情報探索が可能になる，
- 検索結果を 10 件ごとに表示するのではなく，注目する人間に応じて関連する検索結果を表示すれば，検索結果をさまざまな視点で効率良く見ることができる．
- 人間と Web ページの対応付けることから，人間という特徴で Web ページが多重分類されることになり，検索結果をある程度まとめて扱うことができるようになる．
- 人間関係を可視化すると複数のクラスタに分かれることも多いが，クラスタごとに別の内容なので，目的とするクラスタだけを探索するだけでよい．
- 最初に着目した人物の関連する検索結果に目的の情報がない場合には，周囲の人たちを調べれば見つかりやすい．また，関連する内容の情報を調べたい場合には，ある特定の人間の周囲や，その人間が属するクラスタを調べればよい．

3.4 システム構成

本システムの構成図を**図 2** に示す．通常サーチエンジンで使用される全文検索用の転置索引に加えて，HTML ファイルから人名を抽出し，人名検索用の人名索引を作成して使用する．

利用者が検索 GUI に検索語を入力したときには，まず全文検索を行い，そのスコアが高い上位の検索結果の Web ページに出現する人名を検索し，それらの関係を解析して，検索 GUI 部と人間関係可視化部に表示する．

検索結果と人間関係の情報はメモリ上に保持されるので，以後検索 GUI 上で，検索結果に含まれる Web ページ，Web サーバ，人間の関係を用いて関連する情報を対話的に表示したり，人間関係をたどりながら検索結果を閲覧したりすることができる．また，人間関係はネットワーク構造としても可視化されるが，これ

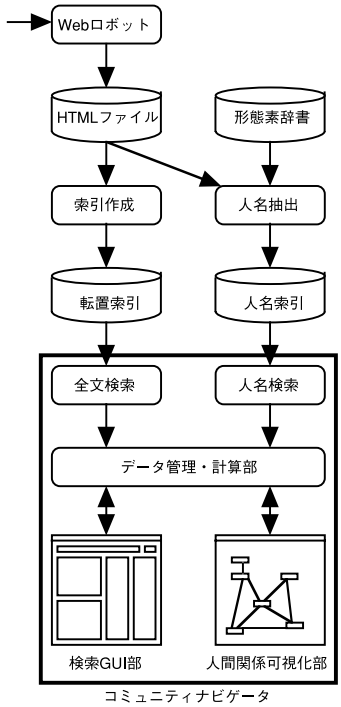

図 2 システム構成図
Fig. 2 System overview.

も検索 GUI と連動しているために，ネットワーク構造上を移動しながら検索結果を閲覧することができる．

4. 人間関係の抽出

ここでは，サーチエンジンの検索結果から人間関係を求める方法について説明する．まず最初に与えられたトピックにより適合する Web ページを検索し，次にそのトピックの分野における主要な人名を検索し，最後に強い関係を持つ人物に着目して人間関係を抽出する．

4.1 人 名 抽 出

情報探索で人間関係を高速に利用できるようにするために，あらかじめ Web ページから人名を抽出し，そのデータベースを作成する．このデータベースでは，各 Web ページに一意に割り当てられた文書 ID をキーとして，その Web ページに登場する人名を検索できる．

まず HTML テキストからタグやコメントを削除した後で，テキスト部分を MeCab [20] で形態素解析し，得られた形態素の並びに対して，姓・名または姓・空白記号・名の連続する並び，そして特殊な人名を抽出する．なお，特殊な人名とは,「ビートたけし」のように，有名人が別名や芸名として使用する人名である．ただし，人名のアルファベット表記には対応していない．一般に姓，名または別名で呼ぶことも多いが，本論文で基本的に姓と名の組に限定するのは，巨大な Web 空間では姓，名，別名だけでは人物を同定できないからである．

なお，日本人の人名をできる限り多くカバーするために，ipadic に大量に人名データを追加し，辞書の人名関係の項目数を約 2.5 倍にし，このときの人名抽出の精度は 0.935，再現率は 0.853 であった[11]．より高度な固有表現抽出技術を用いれば，精度をさらに改善できると思われる．

4.2 Web ページの検索

最初に，指定されたトピックに関連が高い Web ページ群を抽出するために，指定された検索語で Web ページ群を全文検索して，検索された Web 文書のスコアを計算し，スコア順に並べ替える．なお，全文検索には，サーチエンジン ODIN に使用していた全文検索エンジン Jerky を用いている．Jerky では，アンカーテキストとハイパーリンク構造を考慮した検索が可能であり，Google と同様に Web 空間における情報の権威の高さを考慮した順序で並べ替えられている[1]．

なお，ミラーサイトやサイトの移行などの原因により，同じ内容の Web ページが存在した場合には，スコアが低い方を無視する☆．この理由は，まったく同じ内容のページが分散して存在すると，それに含まれる人名を相互に関連づけられることで，人間関係として誤抽出されるからである．

4.3 人名の抽出

検索結果の Web ページから上位 n 件までの Web 文書を選択し，その文書 ID で人名データベースを検索して，人名を Web ページの中の出現順に格納した人名リストを求める．n は，検索結果数の人名の出現率に応じて 100〜5,000 の間を変化させている．上位 n 件に制限することにより，検索語に適合した検索結果だけを解析できること，対象データ量を削減し処理を高速化することのほかに，被リンク数が多い権威ある Web ページに掲載されている人名に限定して抽出できる．

ただし，この手法で得られる人名集合は非常に大規模で，与えられたトピックと関連がないノイズも多く含まれ，そのままでは可視化も難しい．また，検索結果中の人名の順序付けをどうするかという問題も生じる．

そこで本論文では，ある人物があるトピックの Web

☆ 通常の全文検索と異なり，2 つの Web ページがまったく同じ内容であっても，アンカーテキストとリンク関係の違いによって得られるスコアは異なる．

部分空間に対して影響を与える度合いを**影響度**と呼び，これを人名の制限や検索結果の順序付けに使用する．影響度は，次のように定義する．

$$E(p) = |\{s|d \in s \wedge d \in R_{k,n} \cap D_p\}| \quad (1)$$

つまり，ある人物 p の影響度 $E(p)$ は，検索語 k の検索結果の上位 n 件に含まれる検索結果 $R_{k,n}$ 中の，人名 p が出現する文書群 D_p を含むサーバ数である．

たとえば，検索結果の出現数で順序付けすると，ニュースサイトのライタが上位にくることが多いが，これは，その分野に影響を与えているとはいい難く，実際に利用者がライタの存在を記憶していないことも多く，適切ではない．このような問題には，従来のネットワーク構造から求める中心性の定義では対処できない．逆に，著名人は，単一 Web サーバ内の出現頻度が多いとは限らないが，数多くの Web サーバに登場する傾向があるために，影響度の利用は比較的妥当だと思われる．

また，影響度で順序付けするだけでなく，閾値 e よりも小さい影響度を持つ人名を処理から除外する．情報探索に役立つような著名な人物を残しながらも，与えられたトピックにほとんど影響を与えない人物または無関係な人物を削除して人名集合を縮小することで，処理を高速化し，可視化を容易にする．なお，単独の Web サーバにしか人名が現れない人は，その活動が他にほとんど認識されていないと考え，通常は e は 2 以上の値で用いている．

4.4 人間関係の抽出

人々が互いに関係があるかどうかは，人名の**共起**（co-occurence）に基づいて判断する．なお，人名の共起関係は，大きく 2 種類に分類できる．たとえば，論文や書籍の共著，同じシンポジウムにおける発表など，活動の場を共有している場合には，人名が共起する．そうでなくても，同じ特集記事への掲載，リンク集や Web ディレクトリの同じカテゴリなど，その活動が同一文脈で言及されたり，同じカテゴリに属すると判断されたりする場合にも，人名は共起する．前者を**直接的な共起関係**と呼び，後者を**間接的な共起関係**と呼ぶ．

実は，間接的な共起関係の場合には，必ずしも互いに直接な関係があるとは限らない．しかし，同一文脈で言及される場合には，その著者は両者に関係があると見なしている場合が多く，同じカテゴリに属すると判断される場合にも，両者は互いに存在を認識していることが多いと考えられる．そこで，本論文では，間接的な共起関係を直接的な共起関係と区別せずに扱う．

ここで，実際にどのように共起した場合に関係があると見なすかを考える．NEXAS//KeyPerson では，同じ Web ページに人名の共起する場合に関係があると見なしたが，多くの人名が登場する Web 文書の存在により，抽出精度や分離性を低下させる要因になりやすかった．

そこで，本論文では，検索結果の上位 n 件の各 Web ページに対して，その Web ページに出現する人名を本文中に出現する順序で並べた順序付きリストを作成し，このリスト中で距離 d 以内に共起する人名の組に限り関係があると見なす．実際には，人名 A と B について調べる場合には，ある Web ページの人名リストのそれぞれの各 A に対して，それから d 個目離れた位置までに B が存在するかどうかを調べる．これは，人名リスト中で近接している人名の組ほど何らかの関係がある可能性が高く，反対に非常に離れた位置にある人名の組は何も関係がない可能性が高いからである．たとえば，研究会の発表リストを考えた場合には，共著者は互いに強い関係があり，次に同じセッションの発表者も関係するが，別のセッションの発表者とは，必ずしも関係があるとはいえないが，これらは人名リスト内の距離に反映されていると考えられる．

なお，このように人間関係を限定するために，実際に強い関係があるにもかかわらず検出されない可能性が存在する．しかし，該当分野に強い影響を持つ人たちは，互いに一緒に行動する傾向があることから，ある Web ページで必ずしも関係があると見なされなくても，別の Web ページで関係があると見なされると推測される．

5. 情報探索と可視化

5.1 検索結果の表示

Web 空間の人間関係やコミュニティを解析して，情報探索に適用するために，コミュニティナビゲータと呼ぶプログラムを試作した．この実行画面を**図 3** に示す．

上部にある検索フィールドに検索語を入力し，検索結果はその下部の 4 つのリストで表示される．左下のリストは検索結果中で人名を含む Web ページ群，左上のリストはそれらの Web サーバ群，中央のリストは検索結果に含まれる人名，そして右のリストは中央のリストのある人名が選択された場合に，その人名と共起する人名群を表示する．Web サーバの URL の右には，そのドメインの管理組織名と含まれる Web ページ数を，人名の右には影響度を表示している．この組織名の表示により，ある人物がどのようなコミュニティに関係しているかを知ることができる．表示順

図 3 コミュニティナビゲータの実行例
Fig. 3 Community navigator.

は，Web ページがサーチエンジンのスコア順，Web サーバが Web ページ総数順，人名は影響度順である．

たとえば，この実行例では，“XML” で検索し，116,965 件が検索された．$n = 1000$，$e = 2$ の場合には，該当する人名は 50 個，それらの人名を含む検索結果は 97 ページ，それらの人名を含むサーバ数は 41 個である．

5.2 検索 GUI 部

情報探索に関しては，情報の相関関係を表示する機能と，人間関係をたどる機能を提供している．

検索結果の中から相関関係を調べながら目的の情報を探すことができるように，Web ページ，Web サーバ，人名のいずれかを選択すれば，それに関係がある他の項目をリストの上方に集めて反転表示し，さらに選択された人名と共起する人名を右のリストに追加する．Web ページまたは Web サーバが選択された場合には，それに出現する人名は複数になるが，右のリストにはそれらと共起する人名をすべてまとめて表示する．たとえば，図 3 では，人名から 3 番目に高い影響度を持つ「浅海智晴」を選択し，その人名が出現する Web ページ・Web サーバと，その人名と共起する人名群を表示している．$d = 3$ の場合には，出現するサーバ数は 7 個，Web ページ数は 10 ページ，共起する人名は 8 個である．

さらに，人間関係をたどることができるように，右の共起する人名のリストからある人名を選択すると，中央の情報に出現する人名のリストからその人名を選択し，全リストを再表示するようにしている．たとえば，図 3 の右のリストの「山本陽平」を選択すると，中央のリストの「山本陽平」が選択され，右のリストにはその人物に関連する人名が新たに表示される．これを繰り返すことで，人間関係をたどりながら，情報を閲覧することができる．

5.3 人間関係可視化部

情報探索をより容易にするために，このプログラムは人間関係を可視化する機能を提供している．図 3 の検索結果の人物群で影響度が 2 以上の人物に関してばねモデルで可視化した例を，**図 4** に示す．

この人間関係ネットワークのノード数は 42 個，エッジ数は 83 本である．なお，ノード数が検索された人名数より少なくなっているのは，他の人名と共起しない人名は除去しているからである．各ノードの数字は影響度であり，各エッジの数字は共起数である．

情報探索を検索結果画面と人間関係画面は互いに連動し，情報探索に利用できる．検索結果である人名を選択すると，可視化画面でもその選択された人名は濃い色で，その人名と共起する人名は薄い色で反転表示される．逆に，ネットワークの人名を右クリックすれば，検索結果画面の人名が選択される．これにより，リスト表示では難しい概観の把握や，異なる人名クラスタへのジャンプを，より容易にすることができる．

また，可視化することにより，4 つのクラスタに分かれていることが分かるが，これは検索 GUI 部では分からない．ただし，検索 GUI 部のような大規模な人間関係の表示は困難である．

なお，NEXAS//KeyPerson における可視化手法では，人名の共起の度合い，ネットワークの各ノードごとに共起度が高い順に 2 ノードを選択して，それに対するエッジを表示した．これは，与えられた検索語に関連する人間関係を簡潔に表示するためであり，このような制約を適用しないと，大量の人名が登場する Web ページの存在のために人間関係が過度に抽出されやすいからである．

しかし，本論文のように人間関係を情報探索の経路とする場合は，このような局所的な制約は探索経路が極端に減少するために適切ではなく，中心人物も分かりにくくなる問題が生じるので，影響度がある閾値 e を超える人物をすべて表示する．実際にはすでに述べたように，Web ページの人名リスト中で距離 d 以内にある人名だけを共起したと見なすことから，抽出されるエッジ数は高い精度で適切に減少し，より広範囲

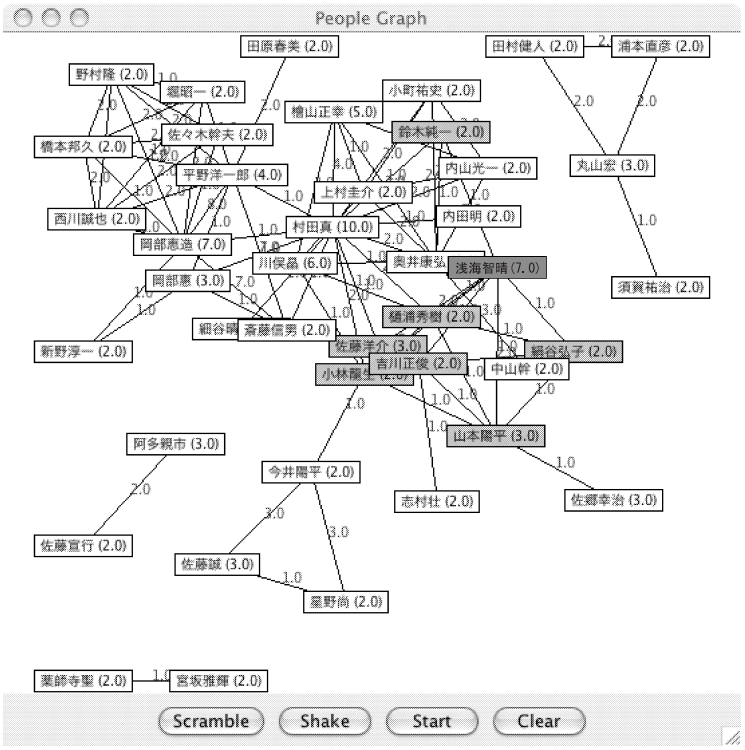

図 4 人間関係の可視化例
Fig. 4 Visualization example of human relationships.

に影響を与える人物が分かりやすくなるとともに，より適切なクラスタが生成されやすくなっている．

6. 評　　価

6.1 評価用 Web データ

評価用に日本語で書かれた Web ページを大量に収集する必要があったが，最近は com ドメインを取得する日本企業が増えたことから，jp ドメイン内だけでは有名・重要な Web サイトが大量に欠落してしまう問題がある．

そこで，jp ドメイン内に加えて，jp ドメインから日本語を含むアンカーテキストでリンクされている Web ページを収集した．これは，アンカーテキストは一般にリンク先の適切な要約であることから，日本の Web ページから日本語を含むアンカーテキストでリンクされていれば，リンク先の内容も日本語である確率が高いという経験則に基づいている．

この収集方法を実装した Web ロボットを用いて，2003 年 7 月に 52,302,804 ページの HTML ファイルを収集して，評価用データとして使用した．

6.2 Web 空間の人名分布の分析

まず，最初に Web 空間における人名の分布について解析する．人名は，13,922,012 ページに出現し，これは全体の 26.6%である．共起解析の対象になる，人名が 2 つ以上含まれている Web ページは，8,716,159 ページであり，これは全体の 16.7%である．この結果から，人名は Web 空間の比較的広い範囲に出現する特徴の 1 つだと考えられる．

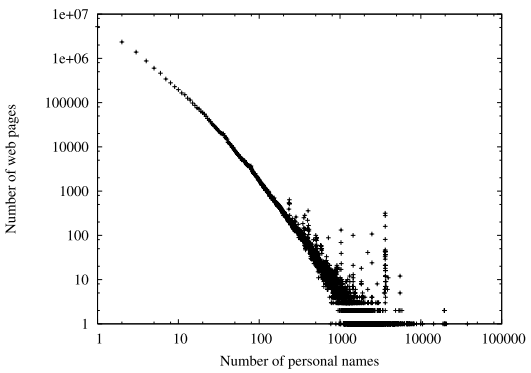

図 5 人名の出現頻度
Fig. 5 Frequency of personal names.

また，**図 5** に，Web ページに登場する人名数と Web ページ数の関係を示す．この分布は冪分布になり，人名が 10 個以上現れる Web ページは 2,194,268 ペー

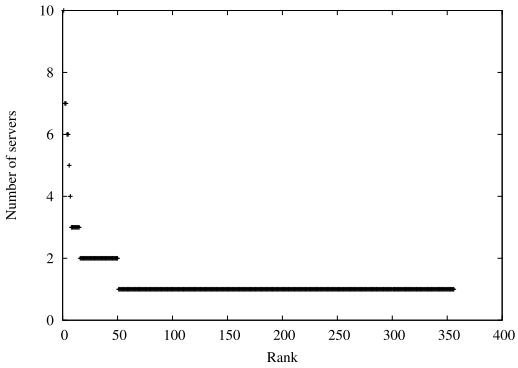

図 6 人名出現サーバ数の分布
Fig. 6 Distribution of the number of web servers on which personal names appear.

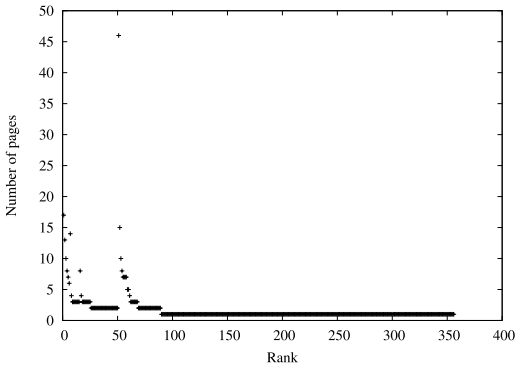

図 7 人名の出現ページ数の分布
Fig. 7 Distribution of the number of web pages on which personal names appear.

ジで全体の 4.2%，100 個以上現れる Web ページは 124,937 ページで全体の 0.24%である．1 ページあたりの平均人数は比較的少ないが，人数がかなり多いページが存在することは，抽出する人間関係の精度が低下する危険を示している．

次に，“XML” で検索したときの，各人名の出現サーバ数と出現ページ数を，まず出現サーバ数でソートしてから，同じ出現サーバ数の人名をさらに出現ページ数でソートして順位付けし，**図 6** に順位と出現サーバ数を，**図 7** に順位と出現ページ数の関係を示す．

図 7 で興味深いのは，出現ページ数のピークは出現サーバ数が多いときではなく 1 のときであることである．出現サーバ数が 1 にもかかわらず，出現ページ数が 46 ページの人物と出現ページ数が 10 ページの人物はどちらもライタであり，出現ページ数 10 ページの「額田王」のように，例題に使用された人名もあった．これは，出現ページ数に頼った手法では情報探索に適さない人物が抽出される危険があることを示している．これに対して図 6 は，サーバ単位にまとめたために要素数が少なすぎることから両対数にはしなかったが，冪分布に近い分布を示している．なお，別の検索語で行った場合も図 6 は同じ傾向を示すが，図 7 に示したような現象は必ず起こるとは限らない．

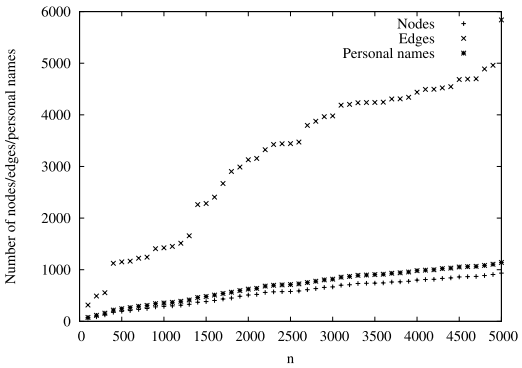

図 8 n にともなうノード数とエッジ数の変化
Fig. 8 Number of nodes and edges with varying n.

6.3 パラメータと人間関係の分析

人間関係を抽出するために使用する n，e，d の 3 つのパラメータが，人間関係の抽出に与える影響を調べる．検索語は，図 3 と同様に “XML” を用いた．

まず，検索結果の n を変化させたときの，人名数，人名の共起ネットワークを構成する人名のノード数とエッジ数の変化を**図 8** に示す．ここで，e が影響しないように $e = 1$ を，d は大量の名前を含む Web ページにだけ影響するように $d = 40$ を用いた．

定義から，Web ページに単独でしか出現しない人名はネットワークに含まれず，ノード数は人名数よりも少なくなる．また，人名数とノード数はほぼ単調増加であるが，エッジ数の増加は複雑であり，急に増加する部分があるのが分かる．これは，多量に人名を含むページが検索結果中に現れた場合に，一気に増加するからと推測できる．

次に，影響度の e を変化させたときの，ノード数とエッジ数の変化を**図 9** に示す．他の値は，n は比較的中間の値として $n = 1000$ を，d は $d = 40$ を用いた．

影響度が 1 の人名は，ある特定のサーバにしか出現しない人名であり，このような人間は情報の探索パスとして適切ではない．これを除き，複数サーバに出現する有名な人名にするだけでも，かなり絞り込むことができるのが分かる．

また，人名間距離の d を変化させたときの，ノード数とエッジ数の変化を**図 10** に示す．他の値は，$n = 1000$，$e = 1$ を用いた．

たとえば，$d = 3$ の場合には，着目した人名の前後に出現する計 6 人が共起と判断されることになる．し

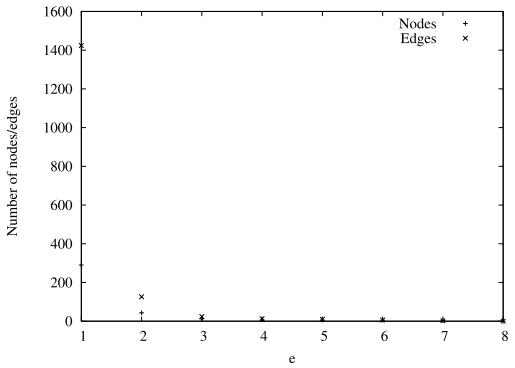

図 9 e にともなうノード数とエッジ数の変化
Fig. 9 Number of nodes and edges with varying e.

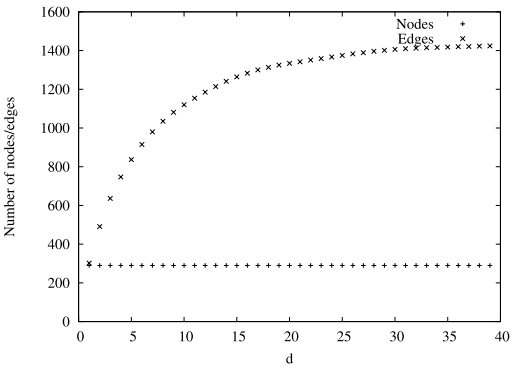

図 10 d にともなうノード数とエッジ数の変化
Fig. 10 Number of nodes and edges with varying d.

かし，d を必要以上に大きくしても，その間に高い確率で直接の関係があると推測することはできないだけでなく，抽出されるエッジ数を大幅に増加させてしまう危険があるのが分かる．

6.4 人物と人間関係の抽出精度

次に，人物と人間関係の抽出精度を評価する．

まず，検索語 “XML”，$n = 1000$，$e = 2$ で得られた 50 人の人物が適切であるかについて評価した．評価には，本プログラムの検索結果の閲覧を主とし，さらにより詳細な情報を得るために Google の検索も併用した．また，判断の指標としては,「XML に関して具体的な活動を行っている人物」とし，ライタなどの XML 関係の活動で世の中に影響を与えていない人物は不可とした．その結果，人名の抽出ミスが 3 件，製品紹介記事で XML に言及したライタが 1 件，関連する分野で活動してはいるが直接の貢献がない人物が 1 件存在し，45 件が適合し，抽出精度は 90%であった．なお，人名の抽出ミスの場合でも該当する人物は適合していた．この結果から，人物の抽出精度は良好であると判断できる．

$E(p) = 1$ である人名の中には，単に XML に関して具体的な活動をしていないだけでなく，まったく無関係な人物も多く存在し，影響度による選別が比較的有効に働いていると推測される．

さらに，$d = 3$ として人名の共起関係を解析し，ノード数 42，エッジ数 83 の人間関係ネットワークを抽出し，抽出されたエッジが人間関係として適切であるかを評価した．この判断の指標としては,「該当する 2 人の人物が実際に活動の場を共有している」とし，たとえば単にあるページで同時に言及されている場合は不可とした．この際に上記の不適合とされたライタは単独で出現するために除外されている．この結果，名前の抽出ミスに関連するエッジが 6 本，上記の直接関係のない人物のエッジが 1 本，別の XML 書籍の著者だが同じページで紹介されていたエッジが 1 本存在し，75 本のエッジが適合し，抽出精度は 90.4%であった．ただし，別の書籍であるが，同じプロジェクトの活動として紹介されていた 1 本は適合とした．また，名前の抽出ミスに関連するエッジでも，内容は適合していた．この結果から，人間関係の抽出精度も良好だと考えられる．

適合とした人間関係は，論文・仕様書・書籍の共著，XML 専門誌への記事の同時掲載,「XML 開発者の日」などの会議の発表,「XML コンソーシアム」,「XML Publishing Forum」,「XML 技術者育成推進委員会」などの団体の運営関係者であり，実世界の具体的なエンティティや活動と密接な関連を持つものが大部分であった．これから，人間関係を利用した情報探索は，Web 空間の情報を手がかりにした実世界の探索の実現につながると考えられる．

6.5 情報探索における人間関係利用の分析

人間関係は現実社会において有効な情報探索経路として使われており，たとえば ReferralWeb も本来は査読者を探すシステムである．しかし，本論文のように得られた人間関係を Web 空間に対応付ける場合の有効性は，明らかではない．そこで最後に，ある検索語に対して得られた人間関係を情報探索という目的に利用する場合の妥当性を分析する．

まず，さまざまな検索語に対して，$n = 1000$，$e = 2$ という同じ条件で検索した結果を，検索結果の少ない順から**表 1** に示す．ページ数とサーバ数は，この条件下で人名が出現する数であり，単に人名が出現する場合（$e = 1$）には図 9 に示されているように，もっと数が多くなる．

この結果から，検索結果数が多いからといって，必ずしも得られる人名数も多いわけではなく，実際にどの

表 1 さまざまな検索語に対する検索結果
Table 1 Search results for various queries.

検索語	検索結果数	ページ数	サーバ数	人名数
ネットワーク分析	611	287	202	1,063
自然言語処理	6,395	671	185	2,173
機械翻訳	6,754	838	156	1,034
データマイニング	11,215	303	194	931
Lisp	19,781	4	4	2
人工知能	29,435	487	214	1,044
blog	43,786	236	62	63
オブジェクト指向	44,026	149	55	153
卓球	103,709	395	47	56
XML	116,965	97	41	50
情報検索	187,558	60	49	45
テニス	406,728	128	54	43
Java	465,869	53	41	22
野球	854,952	305	80	109
サッカー	910,281	471	161	95

程度の規模の人間関係が抽出されるかは検索語に大きく依存することが分かる．そこで，検索結果と人名との間の関連性の判断や情報探索を効率的に行うためには，検索結果と得られた人間関係の規模のバランスがとれている必要があると考えられる．効率的な情報探索という観点からは，比較的多くの検索結果に対して，比較的少ない著名人のネットワークが生成されるのが望ましいと考えている．たとえば，検索語「XML」の場合では，97 ページを 50 個の人名で探索することができ，図 4 から分かるように，その経路も比較的粗であることから，情報探索の効率も良いと考えられる．

これに対して，表 1 で一番人名数が多い検索語「自然言語処理」の場合には，671 ページに対して 2,173 個の人名で探索することになるので探索経路が重複するとともに，人名と検索結果の関係の重なりも多くなってしまって，人名を指定して検索結果を閲覧する利点が損なわれてしまう．$n = 100$ にすれば人名数を 85 個まで数を減らすことはできるが，今度は扱う検索結果が 58 ページ，32 サーバに減少してしまうので，探索範囲が狭くなってしまう．これは，検索語が示すトピックが専門的で，関係者の活動範囲が狭いが，その活動は活発であるからと推測される．そこで，代わりに $e = 9$ とすれば，人名数を 74 個まで減らしながらも，検索結果は 495 ページ，142 サーバになる．この可視化例を**図 11** に示す．このノード数は 74，エッジ数は 823 である．つまり，専門的なトピックで多くの専門家が積極的に活動するような場合には，e を増やして主要な人物に絞り込む方が，情報探索範囲を保ちながら人名を減らす点で適しているのが分かる．しかし，利用者に人名を提示するという点では適切になっても，人間関係のエッジ数がノード数に対して非

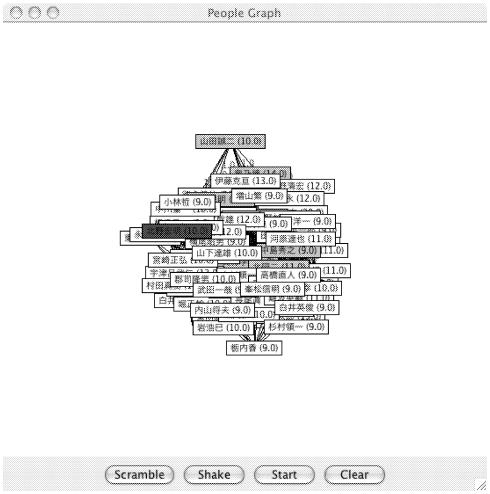

図 11 「自然言語処理」の人間関係の可視化
Fig. 11 Visualization of human relationships for “natural language processing”.

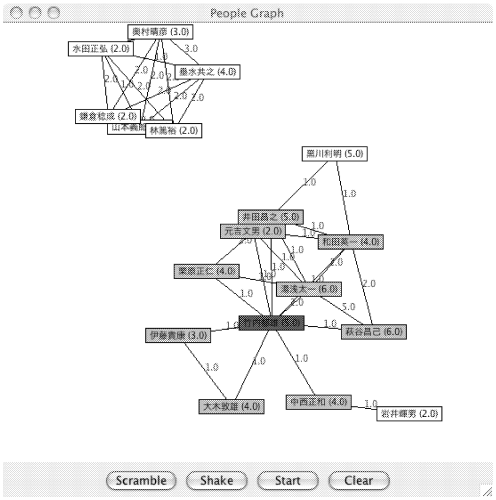

図 12 「Lisp」の人間関係の可視化
Fig. 12 Visualization of human relationships for “lisp”.

常に多いことから，これらの人間関係は密すぎて，情報を探索する経路という点では必ずしも効率的とはいい難い．

逆に，表 1 で一番人名数が少ない検索語「Lisp」の場合には，e をこれ以上減らすのはノイズ除去を考えると難しいので，$n = 3{,}000$ にすれば，81 ページ，35 サーバが検索され，人名数が 27 個まで増加する．この可視化例を**図 12** に示す．このノード数は 18，エッジ数は 35 である．つまり，得られる人名数が少ない場合には，n を増やしてより多くの検索結果を対象にすれば，同様な人間関係が得られることが分かる．ただし，この場合に注意しなければならないのは，検索

結果の上位に人名が現れるようなページが少ない点である．本論文で用いたようなサーチエンジンでは，ハイパーリンク構造から情報の被参照性を解析して検索結果の順位に反映するために，検索結果の上位には利用者の多くが重要だと考えている情報が集まる傾向がある．つまり，検索結果の上位における人名数が非常に少ないということは，人名の登場する Web ページの評価が相対的に低いことを意味しており，さらに検索結果数が少ないことから，そもそも Web における人間活動のアクティビティが低いと考えられる．実は，Lisp というプログラミング言語は 1962 年に登場し，1984 年には Common Lisp として標準化され，現在でもさまざまな分野で利用されてはいるものの，少なくとも国内では新たな研究開発はほとんど行われていない．実際に，図 12 の右下の大きなクラスタは 1980 年代に活躍した人たちであり，現在も積極的に活動しているとはいい難い．このように，すでに安定期または衰退期に入り，人間活動のアクティビティが低下しているような分野のトピックでは，人間関係を使った情報探索手法は，過去ではなく現在 Web 空間で一般に重要だと考えられている情報を探索するためには適していないと考えられる．

7. おわりに

本論文では，人間があるトピックに対して与える影響度と Web ページ内の人名の出現順序を考慮して，検索結果の中の人名の共起を解析することで人間関係を抽出する手法と，抽出された人間関係を情報探索経路として用いる手法，そして人間関係の可視化手法について述べ，処理結果を分析・評価することで有効性を示した．

ハイパー空間の迷子問題を解決するためには，本論文のように実世界のエンティティと関連づけて利用者に提示すること，および実世界のエンティティ間の関係を Web 空間の情報探索経路として提供することは有益だと考えられる．

今後の課題は，処理速度と処理品質のさらなる向上と，より使いやすいユーザインタフェースの開発である．Web 情報処理で対象とするデータは非常に大規模であり，質の向上を考える場合でも処理速度の問題は決して無視できず，これらを両立できる妥当なアルゴリズムであることが必須である．今回の実装では，全文検索，人名検索，および人名・Web ページ・Web サイトデータの相互関係解析などの初期処理を終えた後は対話的にリアルタイムで操作できるが，検索語や使用するパラメータによって得られるネットワーク構造が過度に巨大または緻密になる場合には，初期解析に数十秒以上の大きな遅延が生じてしまうことがあり，その場合の人間関係の効率的な表示も困難である．今後は，人名検索専用データベース作成による高速化，パラメータの自動調節，ネットワークノードとエッジの適切な削減方法などの，さらなる改良が必要である．さらに，今回はリストとネットワークという 2 種類のユーザインタフェースを提示したが，今後は両方の利点を兼ね備えた，よりシンプルなユーザインタフェースを検討する予定である．

謝辞 NEXAS//KeyPerson の開発者であり，本研究に協力していただいた原田昌紀氏と，実世界と Web 空間における情報探索に関する有益な議論に参加していただいた野島久雄氏と新垣紀子氏に感謝する．

参考文献

1) 風間一洋，原田昌紀，佐藤進也：ハイパーリンクとアンカーテキストを利用した情報検索とランキングの一手法，情報処理学会研究会報告 FI-59-3/DD-24-3，pp.17–24 (2000).
2) Silverstein, C., Henzinger, M.R., Marais, H. and Moricz, M.: Analysis of a Very Large AltaVista Query Log, Technical report, Digital SRC (1998).
3) OneStat.com: Most people use 2 word phrases in search engines according to OneStat.com (2004). http://www.onestat.com/html/aboutus_pressbox32.html
4) 神林 隆，清水 奨，佐藤進也，Francis, P.：インターネット情報探索に適したキーワード抽出，情報処理学会研究報告 NL-118-13，pp.79–84 (1996).
5) Kawano, H.: Mondou: Web search engine with textual data mining, *Pacific Rim Conference on Communications, Computers and Signal Processing*, IEEE (1997).
6) 原田昌紀，清水 奨：WWW 検索システムにおける不特定多数の操作履歴の活用，情報処理学会研究報告 OS-74-11/DPS-81-11，pp.61–66 (1997).
7) 戸田浩之，長浜光俊，片岡良治：特徴的な固有表現を用いたラベル指向ナビゲーション手法の提案，情報処理学会研究会報告 DBS-133-12/FI-75-12，pp.99–106 (2004).
8) Kautz, H., Selman, B. and Shah, M.: The Hidden Web, *AI Magazine*, Vol.18, No.2, pp.27–36 (1997).
9) 松尾 豊，友部博教，橋田浩一，中島秀之，石塚満：Web 上の情報からの人間関係ネットワークの抽出，人工知能学会論文誌，Vol.20, No.1, pp.46–56 (2005).
10) Ogata, H., Fukui, T. and Yano, Y.: Social-

PathFinder: Computer Supported Exploration of Social Networks on WWW, *Advanced Research in Computers and Communications in Education*, pp.768–771 (1999).
11) 原田昌紀，佐藤進也，風間一洋：Web 上のキーパーソンの発見と関係の可視化，情報処理学会研究会報告 DBS-130-3/FI-71-3，pp.17–24 (2003).
12) Kautz, H. and Selman, B.: Creating Models of Real-World Communities with ReferralWeb (1998). http://www.cs.washington.edu/homes/kautz/talks/rec98talk.ppt
13) Sherman, C.: Google Alert Automatically Tracks Your Favorite Topics (2004). http://searchenginewatch.com/searchday/article.php/3301451
14) Microsoft: MSN-Harris Interactive Survey Asks: What Is America Searching For? (2004). http://www.microsoft.com/presspass/press/2004/Aug04/08-02SearchPollPR.asp
15) 佐藤進也，風間一洋，福田健介，村上健一郎：実世界指向 Web マイニングによる同姓同名人物の分離，情報処理学会論文誌：データベース，Vol.46, No.SIG 8 (TOD26), pp.26–36 (2005).
16) Granovetter, M.: The Strength of Weak Ties: A Network Theory Revisited, *Sociological Theory*, Vol.1, pp.203–233 (1983).
17) Milgram, S.: The Small World Problem, *Physiology Today*, Vol.2, pp.60–67 (1967).
18) 野島久雄，阪谷　徹：コンピュータネットワーク利用場面における他者の役割，認知科学の発展，日本認知科学会（編），Vol.5, pp.49–71, 講談社 (1992).
19) Barabási, A.-L.：新ネットワーク思考—世界のしくみを読み解く，NHK 出版 (2002). ISBN: 4-14-080743-1.
20) 工藤　拓：MeCab: Yet Another Part-of-Speech and Morphological Analyzer. http://chasen.org/˜taku/software/mecab/

(平成 17 年 3 月 20 日受付)
(平成 17 年 7 月 6 日採録)

（担当編集委員　森本 康彦）

風間　一洋（正会員）

1988 年京都大学大学院工学研究科精密工学専攻修士課程修了．同年日本電信電話（株）入社．現在 NTT 未来ねっと研究所主任研究員および京都大学大学院情報学研究科博士後期課程．分散協調処理，情報検索の研究に従事．ソフトウェア科学会，ACM 各会員．

佐藤　進也（正会員）

1963 年生．1988 年東北大学大学院理学研究科数学専攻修士課程修了．同年日本電信電話（株）入社．協調作業における情報活用支援の研究に従事．現在 NTT 未来ねっと研究所主任研究員．電子情報通信学会，Internet Society，ACM 各会員．

福田　健介

1999 年慶應義塾大学大学院理工学研究科計算機科学専攻後期博士課程修了．同年日本電信電話（株）入社．現在，未来ねっと研究所に所属．この間，2002 年ボストン大学訪問研究員．インターネットトラフィックのダイナミクス，ネットワーク構造の統計的解析等の研究に従事．博士（工学）．ACM 会員．

村上健一郎（正会員）

1955 年生．1979 年九州大学工学部情報工学科卒業．1981 年同大学院修士課程修了．同年日本電信電話公社入社．2005 年法政大学ビジネススクール・イノベーションマネージメント研究科教授，現在に至る．超大型計算機用 OS，記号処理計算機，インターネットパラダイム，超高速インターネットプロトコルの研究に従事．博士（情報科学）．電子情報通信学会，ACM，ソフトウェア科学会各会員．

川上 浩司

1989 年京都大学大学院工学研究科修士課程修了．同年岡山大学工学部情報工学科助手，1998 年京都大学大学院情報学研究科助教授，現在に至る．共生システム設計方法論の研究に従事．人工知能学会等の会員．工学博士．

片井 修

1969 年京都大学工学部機械工学科卒業．同大学院博士課程機械工学科第二専攻修了．1974 年同大学助手．1983 年同助教授．1994 年同教授．1996 年同大学院工学研究科教授．1998 年同大学院情報学研究科教授，現在に至る．その間，1980〜1981 年フランス国 INRIA 客員研究員．主として共生的なシステム構築法に関する研究に従事．工学博士．