2004－ICS－136　(18)
2004／8／5

人名を用いたWeb空間のコミュニティの解析

風間一洋[†]　原田昌紀[†]　佐藤進也[†]
福田健介[†]　川上浩司[††]　片井修[††]

Web の急速な普及に伴い，膨大で多種多様な文書が公開されるようになり，サーチエンジンを用いて Web 空間から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りつつある．半面，その情報量の膨大さ故に，目的とする情報が存在するにも関わらず，利用者がうまく探し出せないことも多く，さらなる情報探索支援方法が期待されている．

本稿では，Web 空間と実世界の情報を結びつける実世界指向の検索を用いることで，利用者の情報探索を容易にすることを試みる．なお，この手がかりとして，Web 文書中に出現する人名に着目する．ある特定の Web サイトの Web 文書集合に人名が登場する場合には，その名前を持つ人物がその Web サイトに何らかの関係があるとみなして，人名や Web サイトが全体に与える影響度を考慮して，ある検索語が Web 空間内で表すトピックに対する人名と Web サイトの 2 部グラフ構造からコミュニティを抽出し，その可視化を試みる．

Web Community Analysis by Personal Names

KAZUHIRO KAZAMA,[†] MASANORI HARADA,[†] SHIN-YA SATO,[†]
KENSUKE FUKUDA,[†] KOJI KAWAKAMI[††] and OSAMU KATAI [††]

With the rapid growth of the web, huge and various documents are opened to public on the web and it is essential for your life and business to search desired information by web search engines. Nevertheless user often fails to search desired information which surely exist on the web because of too many information. So new search assistance methods are desired.

In this paper, we try to assist user's search process by using a real-world oriented searching which combines the web space and the real world. We uses personal names for this purpose. We assume that there is a relation between a person and a web site in the case that his name appears in the web site. And we try to extract communities which belong to the same topic by analyzing the bipartite graph between personal names and web sites in consideration of their effectiveness and visualize them.

1. はじめに

World Wide Web の普及に伴い，膨大で多種多様な情報が公開され，それを仕事や勉強に用いることが日常生活に欠かすことができなくなりつつある．ただし，Web 技術では一元的に情報を管理する手段が提供されていないために，通常は目的の情報を探すためにサーチエンジンを用いる．

しかし，新聞記事や特許データベースと異なり，Web ページは文体や表記法が統一されず，内容の信頼性もさまざまである。さらに、その利用者はわずか 1〜2 語しか用いないために，目指す情報が検索結果中に存在したとしても，膨大な検索結果から望む情報をうまく探し出せないことも多い．

そこで，近年アンカーテキストや PageRank などのリンク解析技術を用いて，検索語に対して一般的に妥当と推測される情報を，検索結果のより上位に配置する技術が開発された．この技術により，たとえば企業名で検索した時にオフィシャルホームページを検索結果の最上位に配置することも可能になった．

しかし，これでは，人間の情報探索過程のほんの一部しか解決されない．たとえば，単なる検索語の出現頻度だけでなく，検索された情報の背景的知識を知り，検索語が持つ複数の意味に従った分類，関連づけ，発展，視点の切り替えをおこなうような情報探索の高度な部分は依然として人間にまかされている．このような状況を改善するための手法の一つとして，Web 空間の情報と実世界のエンティティを結びつけることで，利用者の状況把握や情報空間の探索を容易にする実世界指向検索が考えられる．

† NTT 未来ねっと研究所
NTT Network Innovation Laboratories
†† 京都大学情報学研究科
Graduate School of Informatics, Kyoto University

本稿では，実世界指向検索の要素技術の一つとして，同じトピックに関心をもつ人々の集合である Web コミュニティに注目する．Web 空間を人間の行動や興味に応じてコミュニティとして分類することができれば，その情報を用いてさらに高度な情報探索を支援する可能性が生まれると考えられる．

そこで，ある特定の Web サイトの Web 文書集合に人名が登場する場合には，その名前を持つ人物がその Web サイトに何らかの関係があるとみなして，人名や Web サイトが全体に与える影響度を考慮して，ある検索語が Web 空間内で表すトピックに対する人名と Web サイトの 2 部グラフ構造からコミュニティを抽出し，その可視化を試みる．

2. 関 連 研 究

Web ページの参照関係を解析して Web ページ間の関連性を求めることで，Web コミュニティを発見しようとする研究はいくつか存在する．Kumar らの Web Trawling では，Web ページのアーカイブのハイパーリンクのグラフ構造の中からファンから共参照されることで形成される完全 2 部グラフを探すことで Web コミュニティを抽出する手法を提案した[1]．村田は，完全 2 部グラフ構造を，与えられた数個の Web ページからサーチエンジンのハイパーリンクを逆に辿る機能を用いて漸次的・部分的に求める手法を提案した[2]．Reddy らは，密な 2 部グラフ構造 (DBG: Dense Bipartite Graph) を抽出することで Web コミュニティを抽出する手法を提案した[3]．豊田は，Kleinberg の HITS アルゴリズム[4] を改良してハブとオーソリティの関係から Web コミュニティを発見する手法を提案している[5]．本稿のアプローチも 2 部グラフ構造に着目するが，それはファンまたはハブを作成した情報利用者側の視点ではなく情報作成者の視点に基づくこと，そして Web コミュニティを従来のように Web ページではなく人間から求めることが異なる☆．

情報から人名，組織名，地名，製品名のような固有表現を抽出することを，**固有表現抽出**(Named Entity Extraction) と呼ぶ．固有表現抽出は，質問応答システム[6]，ラベル指向の情報ナビゲーション[7] などに用いられている．我々は，抽出された固有表現をそのまま使用するのではなく，それを元に社会ネットワーク解析を試みる．

Web 空間の社会ネットワーク構造の解析に関する研究もいくつか存在する．Web ページ集合における人名の共出現関係を，Shah の Referral Web ではサーチエンジンを，また Ogata らの SocialPathFinder は Web ロボットを用いて，漸次的・部分的に抽出することで，その社会ネットワーク構造を得る[8),9)]．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 ページ集合を抽出し，各ページに出現する人名集合を求めた後で，その共出現関係を解析して社会ネットワーク構造を求める[10)]．本稿では，後者の手法に基づくが，この方が高速であるだけでなく，Web 空間を部分的ではなく全体的に解析できるので，互いに直接の関連がない複数の Web コミュニティを求めることができる．

なお，Web 空間における社会ネットワークをメタデータとして記述するための試みとして，FOAF が存在する[11)]．FOAF では，電子メールアドレスをネットワーク上における個人の識別子と考え，さまざまな関係を記述する XML/RDF の名前空間を提案している．これに対して，本稿のアプローチは発見的な手法であるために膨大な情報から自動抽出できるのでより広範囲の情報が得られる利点があるが，人名がどのような文脈で使用されているかの判別や同姓同名などの問題や，多彩な属性の抽出が困難という欠点があり，将来的に互いの特徴を生かした利用が望ましいと考えられる．

3. 人名と Web コミュニティ

3.1 人名の利用

本稿では、実世界指向の情報検索を実現するために，Web 空間における人間の活動に注目する．たとえば，注目している分野の人が，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達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だろう．また，ある分野に関係している人の名前や人数は，その分野が一般的又は特殊なのか，また注目されているかどうかなどの判断の基準となるであろう．特に，著名人や専門家に関しては，そ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だろう．つまり，単に Web ページの情報を閲覧するだけでなく，同時に関連する人間の行動が把握できれば，情報の選択や高度な理解に役立てることができると推測される．

しかし，Web 空間内で各人がどのような行動をしているかを観測することは容易ではない．Web 空間の中

☆ この違いゆえに，本稿における Web コミュニティという言葉の定義は若干異なる

では，現実世界の人物は，その名前，つまり人名で参照されることから，本稿では，Web ページから人名を抽出して解析することで，人間の行動を観測する手法を提案する．たとえば，サーチエンジンで人名や別名を使って検索する行為は**エゴサーチ**(ego search) と呼ばれ，自分あるいは他人の行動を知るための重要な情報検索手法としてよく知られている．本稿で提唱するように複数の人間の行動を観察することは，今後サーチエンジンの検索の重要な使用法になる可能性があると考えている．

なお，人間やその集団を表す固有表現としては，他に電子メールアドレス，企業・団体名なども存在する．たとえば，FOAF では個人の識別子として電子メールアドレスを基本としている．本稿で特に人名を使用するのは，電子メールアドレスは個人の識別には有効であったとしても，利用者がそれから実世界の人間を想起できるとは限らないからであり，または企業・団体の活動については，Web コミュニティの節で述べるように，人名の分布を解析することでも得られると考えられるからである．

3.2 人名の性質

我々が人名を使用するにあたって，特に注目した性質は，普遍性，偏在性と共出現性である．

まず，人名は，一般に　Web 空間の中で広範囲に渡って出現する．実際に Web ロボットで収集した約 4000 万ページを分析すると，23.3%が姓と名の組からなる人名を含んでいた[10]．このように比較的広範囲に出現するために，Web ページの普遍的な性質として扱うことができる．

次に，個々の人名に着目すると，Web 空間の中でかなり偏って出現する傾向が強い．実際に一般名詞と人名に関してサーバごとの出現頻度を調べると，人名は普通名詞よりもかなり出現頻度が低くなる傾向が見られる[12]．たとえば，一人の人間の行動範囲は，情報検索時に想定される分野 (例えば,「機械学習」という検索語が示す分野) の一部に過ぎないだけでなく，ある特定の団体 (例えば，学会など) に限定されることもある．また，交際範囲を選択することもある．つまり，偏在性は人間の意志に基づくものであり，逆に偏在性に着目することで膨大な情報を妥当な観点で絞り込める可能性がある．

さらに，書籍執筆，講演会，学会，競技会など，人間は複数で行動することも多い．当然，同じ Web ページ，さらに同じ文章に異なる人名が登場することも多くなるので，逆にそれらの共出現関係を調べれば，論文の共引用分析[13] のように，人間のネットワーク構造を得ることができる．

3.3 Web コミュニティ

本稿では，Web コミュニティを，同じトピックに関心をもち，かつ互いに何らかの関係がある人々の集合と定義する．

人々が同じトピックに関心を持つかどうかは，そのトピックを表すキーワードの共有の有無に基づいて判断する．実際には，与えられたキーワードで全文検索することで，情報空間を絞り込む．なお，キーワードとして人名が与えられた時には，トピックではなく，該当する人物の関連する Web コミュニティを求めることになり，処理そのものは同じでも，その意味が異なってくることに注意されたい．この場合には，後述するような同姓同名問題が発生しやすくなる問題が存在する．

人々が互いに関係があるかどうかは，人名の**共出現**(co-occurence) に基づいて判断する．なお，人名の共出現関係は，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，人名も共出現する．この関係を，**直接的な共出現関係**と呼ぶ．他に，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る場合も，人名は共出現する．この関係を，**間接的な共出現関係**と呼ぶ．

実は，間接的な共出現関係の場合には，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み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，本稿では，間接的な共出現関係を直接的な共出現関係と区別せずに扱う．

ここで，実際にどのように共出現した場合に関係があると見なすかを考える．たとえば，NEXAS//KeyPerson では，同じ Web ページに人名が共出現した場合に，関係があると見なした．この場合は，人名と Web ページは 2 部グラフを形成し，Web ページを介してどのように人間が結びついているかを求めることになり，得られた人間のクラスタも Web コミュニティと見なすことができる．

本稿では，人間の活動の場に注目し，それらが人間によってどのように結びついているかを分析する．たとえば，現実世界の人間の活動の場としては，学会，オープンソース団体，学校，企業などが挙げられるが，

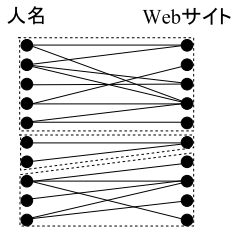

図 1 人名と Web サイトの 2 部グラフの例

現在では，それらの多くが Web 空間に独自の Web サイトを持つ．そこで，Web サイトを活動の場の最小単位と考えて，人名と Web サイトの 2 部グラフを求めて，Web サイトが人間によってどのように結びついているかを解析する．この 2 部グラフの簡単な例を，**図 1** に示す．

この例では，全体に 3 つのクラスタに分類されているが，本稿ではそれぞれのクラスタを 1 つの Web コミュニティとみなす．定義からわかるように，この Web コミュニティには，人名と Web サイトの 2 つの視点があることに注意して頂きたい．

また，現時点では，処理の簡略化のために厳密な Web サイト抽出は試みずに，Web サーバ単位で処理している．これにより，複数の Web サイトが識別できなくなる問題が考えられるが，現実にはオフィシャルサーバを持っている団体も多いこと，また特定のトピックで情報を絞り込むために重複が少なくなることから，深刻な問題にはなっていない．

4. Web コミュニティの抽出

では，実際にどのように Web コミュニティを抽出するかについて説明する．

4.1 人名の抽出

人名を検索できるようにするために，Web ページの文書 ID から，その Web ページに登場する人名を検索するための索引を事前に作成する．

現時点では，人名を抽出するために特別な固有表現抽出技術は用いない．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 で日本語形態素解析し，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．ただし，人名のアルファベット表記には対応していない．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．

人間は，姓だけ，名だけ，又は別名で呼ばれることも多く，Web でも，そのような呼称は多くみられる．しかし，基本的に姓と名の組に限定しているのは，姓，名，または別名だけが用いられるのは，対象範囲が狭い範囲に限定される場合に限られるからであり，Web 空間では，姓，名，別名だけで個人を同定することは難しいからである．

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，約 2.5 倍にしている．この時の人名抽出の精度は 0.935，再現率は 0.853 である[10]．より高度な固有表現抽出技術を用いれば，精度．再現率共に改善することができると思われる．

4.2 Web ページの検索

Web コミュニティを解析するために，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．まず，トピックを限定するために，指定された検索語で Web ページ群を全文検索する．次に，検索された Web 文書のスコアを計算し，スコア順に並び替える．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 空間における情報としての権威の高さを考慮した順序で並び替えられているとみなすことができる[14]．

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，同じページが分散して存在すると，閉じた関係が成立してしまうからである．

4.3 人名の検索

次に，検索結果の Web ページから上位 *n* 件の Web 文書を選択し，その文書 ID から，Web ページに出現する人名を求める．現時点では，*n* は，検索結果数の人名の出現率に応じて 100〜3000 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他に，被リンク数が多い Web ページに掲載されている権威ある人名を優先して抽出できる利点がある．

4.4 Web サイトと人名の関係の解析

次に，得られた検索結果の Web ページを Web サーバ単位でグループ化し，これを Web サイトとして扱う．*m* 個の Web サイトと *n* 個の人名が得られた場合に，この相関関係を，次のように *m* 行 *n* 列の**接続行列**(incidence matrix) で表す．

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

行列 A は Web サイトと人名の 2 部グラフの行列表現であり，成分 a_{ij} は，m 番目の Web サイトに n 番目の人名が出現するどうかを表し，0 か 1 の値を取る．

Web サイト間の関係を表す隣接行列 R_s は，次のように求める．

$$R_s = A \times {}^tA$$

R_s は，m 行 m 列の正方行列である．成分 r_{ij} は，i 番目と j 番目の両方の Web サイトに出現する人名の数である．

また，人名間の関係を表す**隣接行列**(adjacency matrix)R_p は，次のように求める．

$$R_p = {}^tA \times A$$

R_p は，n 行 n 列の正方行列である．成分 r_{ij} は，i 番目と j 番目の両方の人名が出現する Web サイトの数である．

4.5 影 響 度

ただし，この手法で得られる Web サイトと人名の関係は非常に複雑で大規模であり，そのまま可視化するのは難しく，人間が理解するのも困難である．そこで，本稿では，Web サイトの関係を解析する場合には，より広範囲に影響を与える Web サイトだけを抽出し，人間の関係を解析する場合には，より広範囲に影響を与える人間だけを抽出する手法を提唱する．本稿では，この影響を与える度合いを**影響度**と呼ぶ．

なお，より広範囲に影響を与える Web サイトとは，より多くの人間が参加するような Web サイトだと考えられる．たとえば，小規模なグループよりも，大規模な団体の方が，コミュニティ全体に与える影響が大きくなると考えるのは妥当である．また，より広範囲に影響を与える人間とは，より多くの Web サイトに登場するような人間だと考えられる．たとえば，ニュースサイトのライターは検索結果中の出現頻度が比較的高いことが多いが，他の Web サイトにまったく登場しないことも多く，この場合はコミュニティに影響を与えているとは言い難く，実際に利用者がライターの存在を記憶していないことも多い．逆に，著名人は，単一 Web サイトの出現頻度が多いとは限らないが，数多くの Web サイトに登場する傾向がある．

そこで，行列 A における k 番目の Web サイトの影響度 $E_S(k)$ を，次のように定義する．

$$E_S(k) = \sum_{j=1}^{n} a_{kj}$$

すなわち $E_S(k)$ は，k 番目の Web サイトに出現する人名の数である．本稿では，行列 A に対して，$E_S(k)$ が閾値 T_S 以下の Web サイトの列を削除し，同時に削除された Web サイトにしか出現しない人名の列を削除することで，より小さな行数・列数をもつ行列 B を求め，それを Web サイトの隣接行列を求めるために使用する．

さらに，行列 A における k 番目の人名の影響度 $E_P(k)$ を，次のように定義する．

$$E_P(k) = \sum_{i=1}^{m} a_{ik}$$

すなわち $E_P(k)$ は，k 番目の人名が出現する Web サイトの数である．本稿では，行列 A に対して，$E_P(k)$ が閾値 T_P 以下の人名の行を削除し，同時に削除された人名しか出現しない Web サイトの行を削除することで，より小さな行数・列数をもつ行列 C を求め，それを人名の隣接行列を求めるために使用する．

影響度を用いてより小さな接続行列を求める利点は，可視化時にノードから出るエッジの本数を制限するような局所的な制限ではなく，全体的な制限を課すことができるために，より広範囲に影響を与える Web サイトや人間がわかりやすくなることである．

5. Web コミュニティの可視化

Web サイトおよび人名の隣接行列から得られる Web コミュニティを可視化する手法について説明する．

5.1 コミュニティナビゲータ

実験データとしては，Web ロボットを使用して，2001 年 12 月に JP ドメインとそれから 1 ホップで到達できる日本語のサイトを約 4,300 万ページを収集して使用している．この Web 文書集合から抽出されたユニークな人名の数は，約 424 万個である．

このデータを用いて，Web 空間の人間関係やコミュニティを解析するために，コミュニティナビゲータと呼ぶ Web コミュニティを分析するためのプログラムを作成した．このプログラムの画面を**図 2** に示す．画面の最上部に検索語入力フィールドがあり，その下に検索結果，サーバ名，人名を表示するリストが設置されている．プログラム内部では，検索結果，サーバ群，人名の配列と，その対応関係を保持しており，検索結果，サーバ名，人名のリストのどれかの項目をクリックすれば，他のリストの対応する項目が反転表示される．

5.2 可 視 化 例

実際に，Web サイトに関する Web コミュニティの

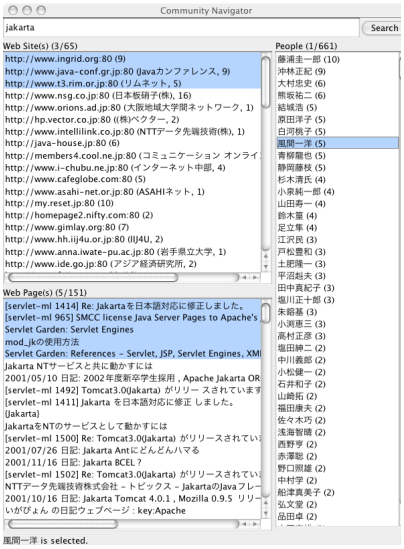

図 2 コミュニティナビゲータ

可視化例を**図 3** に，人名に関する Web コミュニティの可視化例を**図 4** に示す．どちらの表示も，ばねモデルを使用している．

この例の検索語は “jakarta” であり，使用した検索結果は上位 1,000 件であり，影響度の閾値 $T_S = 2$ と $T_P = 2$ より小さい Web サイトと人名を除去した．この場合に，人名の出現する Web ページは 151 ページであり，Web サイトは 65 サイト，登場する人名は 661 人であった．

さらにグラフの画面はコミュニティナビゲータの主画面と連動しており，図 3 および図 4 のグラフのノードをクリックすることで，主画面の該当する項目を反転表示することができる．

どちらの可視化例も 4 つの Web コミュニティが得られており，どちらも左上，右上，左下，右下がそれぞれ，オープンソースの Jakarta プロジェクトに関するコミュニティ，インドネシアの政治に関するコミュニティ，インドネシアの言語に関するコミュニティ，インドネシア外交に関する日本政府のコミュニティである．

隣接行列を求める手法から容易に推測がつくように，Web サイトに関するコミュニティと人名に関するコミュニティは表裏一体で，互いに密接な関係にある．しかし，どのようなコミュニティが得られるかは影響度の閾値 T_S と T_P に強く影響されるので，値によっては必ずしも同数のコミュニティが得られるとは限らないことに注意されたい．これは，人名と Web サイトを求める際に，影響度の閾値 T_S と T_P をそれぞれ適用した 2 つの異なる隣接行列を使用するからである．影響度の閾値を同時に適用すれば，これは回避できるが，実際に実行すると得られる結果が大幅に減少してしまったために，現在のようにしている．

なお，図 4 と NEXAS//KeyPerson の表示との顕著な違いは，本手法の場合には，より広範囲に影響を与える Web サイトや人名が判別できること，およびその定義から人名のネットワーク構造は完全グラフにより近い傾向が見受けられることである．

6. 今後の課題

6.1 Web コミュニティの分離性の向上

可視化の際にしばしば発見される問題は，本来分離されるのが適切だと思われる複数の Web コミュニティが融合してしまうことである．これは，人名に原因がある場合，Web サイトに原因がある場合，Web ページに原因がある場合の 3 つに大きく分類できる．

人名に原因がある場合は，同姓同名の人間が存在する場合である．つまり，同じ名前をもつ人が区別できないために，本来異なるはずのコミュニティがその人名を介して融合してしまう．実際には，検索語で全文検索して指定されたトピックに絞り込むために，同姓同名の出現確率は低くなるために，顕在化するとは限らない．ただし，人名で検索した場合には顕著に出てしまう．そこで，Web 空間の中の人間の活動の場に着目することで分離できないかを検討中である[12]．

Web サイトに原因がある場合は，本稿で用いた Web サーバを Web サイトと見なすという簡略化した定義に原因がある場合である．つまり，巨大な Web サーバ上に，検索語で指定されたトピックであるが，まったく異なる人や団体が管理している Web サイトが共存すると分離できない．実際には，本稿で対象とするように姓と名で構成される人名をわざわざ用いるのは，比較的公式なサイトに集中し，さらに全文検索におけるリンク解析の利用により権威ある Web サイトが選択される傾向があるので，比較的顕在化しにくいが，今後は別の Web サイトの定義も検討したい．

Web ページに原因がある場合は，巨大な名簿にさまざまな人名が列挙されたり，さまざまな話題が同一ページで述べられる場合である．これは共出現したとみなす範囲を制限することで緩和できると推測されるが，計算コストが大幅に増大するために実現が困難である．

6.2 パラメータの自動調整

現在，使用する検索結果数，および影響度の閾値 T_S と T_P は使用者が設定しているが，パラメータの自動調節を検討したい．

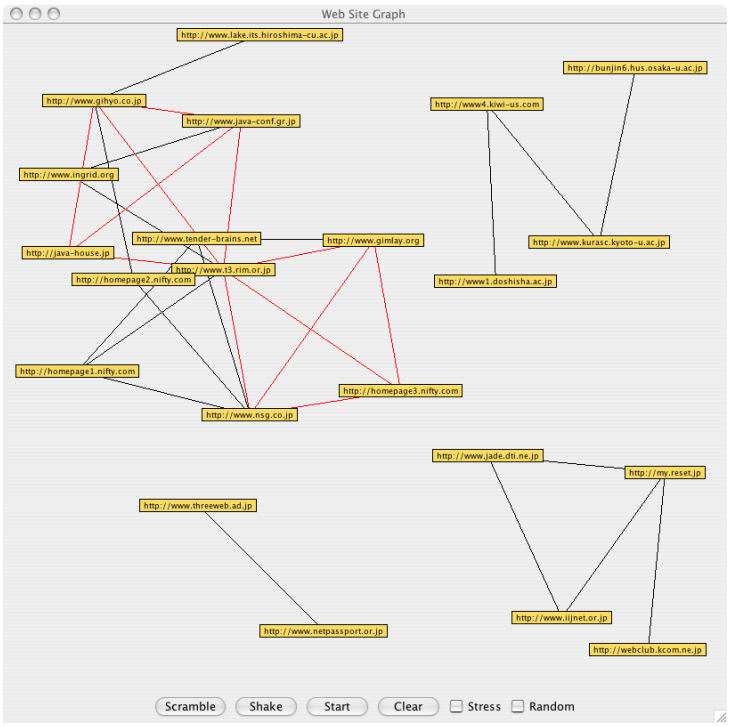

図 3　Web コミュニティの可視化例 (1)

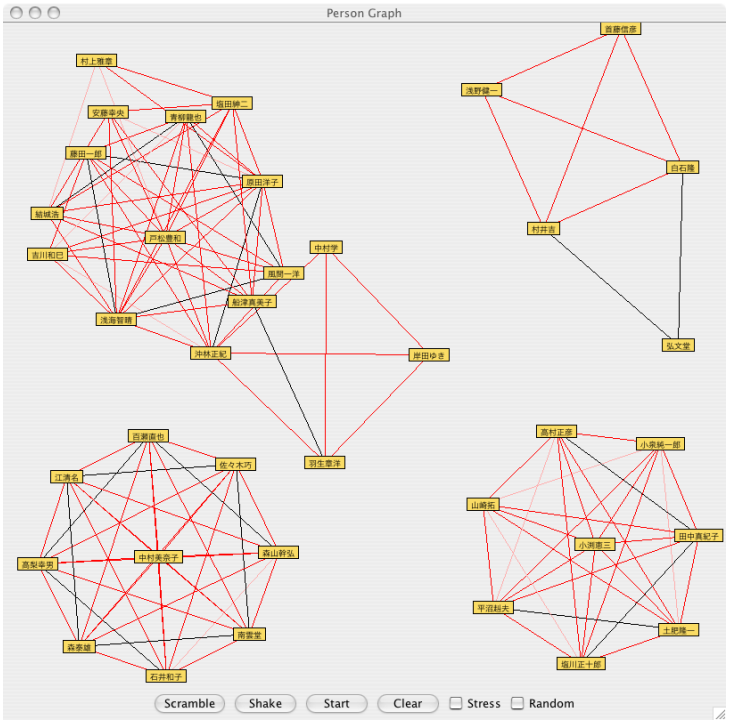

図 4　Web コミュニティの可視化例 (2)

6.3　グラフの描画の改善

現在，グラフの描画にばねモデルを用いているが，これが最前とは言い難く，人手でレイアウトを調節する必要もある．より大規模なグラフ構造の描画とレイ

アウトが適切におこなわれるアルゴリズムを検討したい．

7. さいごに

本稿では，Web 空間と実世界の情報を結びつける実世界指向検索を実現するための要素技術の 1 つとして，人名を手がかりとした Web 空間のコミュニティの解析を試み，得られた結果を影響度を考慮して可視化することにより，異なる Web コミュニティが分離して表示される実例を示した．

本手法は非常に単純であり改善の余地も多いが，Web サイトの情報が人間というエンティティを介してどのように結びつけられているかを利用者に提示することにより，Web 空間を実世界のエンティティを手がかりに探索できる可能性を明らかにしたと考えられる．

今後は本手法の改善とともに，本手法の特性の分析や定量的な評価をおこなう予定である．

参 考 文 献

1) Kumar, R., Raghavan, P., Rajagopalan, S. and Tomkins, A.: Trawling the Web for emerging cyber-communities, *Computer Networks (Amsterdam, Netherlands: 1999)*, Vol. 31, No. 11–16, pp. 1481–1493 (1999).
2) 村田剛志: 参照の共起性に基づく Web コミュニティの発見, 人工知能学会論文誌, Vol. 16, No. 3, pp. 322–329 (2001).
3) Reddy, P. K. and Kitsuregawa, M.: An approach to relate the web communities through bipartite graphs, *Proceedings of the 2nd International Conference on Web Information Systems Engineering*, IEEE (2001).
4) Kleinberg, J. M.: Authoritative sources in a hyperlinked environment, *Journal of the ACM*, Vol. 46, No. 5, pp. 604–632 (1999).
5) 豊田正史: WWW における関連コミュニティ群の発見, 情報処理学会研究会報告 DBS-122, 情報処理学会 (2000).
6) 佐々木裕, 磯崎秀樹, 平博順, 平尾努, 賀沢秀人, 鈴木潤, 国領弘治, 前田英作: SAIQA：大量文書に基づく質問応答システム, 情報処理学会研究会報告 FI-64-12/NL-145-12, 情報処理学会, pp. 77–82 (2001).
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, 情報処理学会, pp. 99–106 (2004).
8) Kautz, H., Selman, B. and Shah, M.: Referral Web: combining social networks and collaborative filtering, *Communications of the ACM*, Vol. 40, No. 3, pp. 63–65 (1997).
9) 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).
10) 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, 情報処理学会 (2003).
11) Brickley, D. and Miller, L.: FOAF Vocabulary Specification, http://xmlns.com/foaf/0.1/ (2004).
12) 佐藤進也, 原田昌紀, 風間一洋: Web 上の「活動の場」に着目した人物の特徴付け, 情報処理学会研究会報告 DBS-133-9/FI-75-9, 情報処理学会 (2004).
13) Small, H.: Co-citation in scientific literature: A new measure of the relationship between two documents, *In Journal of the American Society for Information Science*, pp. 265–269 (1973).
14) 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, 情報処理学会 (2000).