論 文 Technical Papers

Web上の情報からの人間関係ネットワークの抽出

Social Network Extraction from the Web information

松尾 豊 Yutaka Matsuo
独立行政法人 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology
y.matsuo@carc.aist.go.jp, http://www.carc.aist.go.jp/~y.matsuo/

友部 博教 Hironori Tomobe
名古屋大学情報科学研究科
Graduate Scholol of Information Science, Nagoya University
tomobe@nagao.nuie.nagoya-u.ac.jp, http://www.nagao.nuie.nagoya-u.ac.jp/members/tomobe.xml

橋田 浩一 Kôiti Hasida
独立行政法人 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology
hasida.k@carc.aist.go.jp, http://www.carc.aist.go.jp/~hasida.k/

中島 秀之 Hideyuki Nakashima
公立はこだて未来大学
Future University - Hakodate
n.nakashima@fun.ac.jp, http://www.carc.aist.go.jp/~n.nakashima/

石塚 満 Mitsuru Ishizuka
東京大学大学院 情報理工学系研究科
Graduate Scholol of Information Science and Technology, The University of Tokyo
ishizuka@miv.t.u-tokyo.ac.jp, http://www.miv.t.u-tokyo.ac.jp/~ishizuka/

keywords: social network, search engine, Web mining

Summary

Social relation plays an important role in a real community. This paper proposes a new approach to automatically obtain a social network, especially a collaboration network of researchers, of a community from the Web: Nodes are given beforehand. Edges are added consulting to a Web search engine. If two names co-occurs in a lot of Web documents, we assume these two have a strong relation. Moreover, by analyzing the retrieved documents, edge labels are assigned to edges to represent classes of relations such as co-author, same laboratory, same project, or same conference. We operated our system at JSAI2003. Various evaluations are made to show the effectiveness of our approach.

1. ま え が き

「行為を決定するのは，行為者を取り囲む関係構造である.」これが，社会学におけるネットワーク分析の基本的な考え方である [安田 97]．この考え方は，行為者の行為を決定するのは行為者個人であり，したがって行為者の属性によって分析することができるという，いわば「属性主義」とは対極にある．実際には，個人の行為は，属性と関係構造の双方にその要因を求められるだろうが，個人の行動を理解，学習，予測し，個人にとって価値のある情報提示を目指す情報支援の研究において，個人の属性だけでなく，個人を取り囲む関係構造に着目することは重要な方向性である．

近年，多くのコンピュータやセンサが環境や機器に埋め込まれ，多様な情報通信インフラがシームレスに接続されるユビキタスネットワークを実現するための研究が行われている．このような環境下でユーザの状況に応じた情報支援を行うには，ユーザの文脈を適切に推定する技術が必要である [中島 01]．ユーザの文脈を理解するには，ユーザに付随する属性情報だけではなく，ユーザを取り巻く他者との関係構造を把握することも重要である．

また，現在，Web ページや論文などの文書を対象とした多くの情報検索システムがある．しかし，[Kautz 97] で述べられているように，本当に価値がある情報は公開されておらず，人に直接聞かなければいけないことも多い．そのため，目的とする人と自分をつなぐ関係性のパスを発見することは重要である．実際，共同研究を行う，査読をお願いする，学会を運営するなど，仕事・研究の場面において，人間関係が少なからず活用されることも多い．“Networking” という言葉で表されるように，いかにネットワークを作って自分の活動を効率的に行う環境を整えていくかは，個々の研究者にとって，またひとつの学問分野全体においても重要な視点である．

しかし，個人を取り巻くネットワークに着目した情報支援の研究は，これまで少なかった．それは，個人のネットワークに関する情報を取得することが困難であるためである．社会学の分野で，個人のもつネットワークを抽出するためによく用いられる方法は，ネットワーク・クエスチョンとよばれる方法である [安田 97]．例えば,「過去半年のあいだに，あなたにとって重要なことを話し合っ

た人々は誰でしたか」という質問を行うことで個人の持つネットワークや関係性などを明らかにしていく．しかし，このネットワーク・クエスチョンを多くの人に定期的に行うのは難しい．近年，セマンティックウェブの文脈の中で，FOAF[FOAF] という知人関係を記述するメタデータの形式も提案されている．しかし，FOAF は基本的に，ユーザが知人を記述したデータを各自のサーバに置いておくものであり，知人関係を記述する手間の解決にはなっていない．こういったネットワークを，インタラクションのデータから自動的に取得しようという試みもある．例えば，ブックマークを利用して関係性を得る [濱崎 02]，E メールのやりとりから関係を得る [Tyler 03] などの研究である．しかし，こういった研究はプライバシーに関わる情報源から関係性を得る方法であるため，どうしても閉じたコミュニティ，限定した期間で行わざるを得ず，大規模でオープンなコミュニティに対して適用することは難しい．

一方で，近年，Web 上にある多様で大量の情報から，隠された構造や重要な情報を見つけ出す Web マイニングの研究が盛んに行われている．Web のリンク関係から重要なページを発見したり [Brin 98]，あるトピックに関する Web 上のコミュニティ [Kleinberg 98][Kumar 99] を発見する，特定の 2 人の人間をつなぐ知り合い関係のパスを抽出したり [Kautz 97]，参照の共起性からコミュニティを発見する [村田 01]，またあるページの評判情報を抽出する研究 [Rafiei 00] など，さまざまな研究が行われている．

本論文では，Web マイニングの技術を用いて，Web 上の情報だけから特定のコミュニティ（人工知能学会）の人間関係を自動的に抽出する手法を提案する．検索エンジンを利用して，特定の 2 者に関係する文書を検索し，そのヒット件数や文書の内容から，関係の強さおよび関係の種類を判断する．人をノード，関係をエッジ，関係の種類をエッジのラベルとしたネットワークを本論文では**人間関係ネットワーク**と呼ぶ．人間関係というと，公的な関係から私的な関係までさまざまなものが含まれるが，本論文で意図する人間関係は，「研究者間の協働関係」に限定している．

我々は，第 17 回 人工知能学会全国大会（JSAI2003）において，イベント空間情報支援 [西村 04] におけるシステムのひとつとして，人間関係ネットワークの表示システムを運用した．会場での参加者間のコミュニケーションに役立ててもらう目的で，位置表示システム，スケジューリング支援システムなどとも連携したサービスを提供した．本論文中では，この JSAI2003 の具体例を出しながら説明を行っていくことにする．

本論文ではまず 2 章で人間関係ネットワークの利用法とそのアプリケーションについて述べる．3 章では Web 上から人間関係ネットワークを抽出する方法について述べる．4 章では人間関係ネットワークの実例を紹介し，表示法について述べる．そして 5 章では人間関係ネットワークの分析と評価を示し，6 章で議論と関連研究の紹介を行う．

2. ノードとエッジの抽出

2·1 基本的なアルゴリズム

人間関係ネットワークを構成するメンバーはあらかじめ決められているとする．つまり，ネットワークのノードは所与である．例えば，JSAI2003 などの学会の参加者[*1]の氏名は，開催に先だって公開されている．また特定の学会誌の過去の論文の著者リスト，情報系の研究者のリストなどを入手すれば，特定の学会や分野における研究者の氏名リストを入手することが可能である[*2]．JSAI2003 におけるネットワークの場合には，1999 年から 2003 年まで 5 年間の人工知能学会全国大会における著者および共著者をノードとした．後述するように，同姓同名の問題に対処するために，参加者の氏名に加え所属情報も得る．ただし，我々が個人に関する情報として事前に用意するのは，氏名と所属だけである．

次に，ノード間にエッジを付与する処理を行う．基本的なアルゴリズムは非常にシンプルである．例えば，「松尾豊」と「石塚満」の関係を調べるときには，検索エンジンに

“松尾豊　石塚満”

と入力する．(両者は AND の関係である．)「松尾豊 AND 石塚満」の場合には，156 件のヒットがあるのに対し[*3]，「松尾豊 AND 溝口理一郎」の場合には 7 件のヒットしかない．「石塚満」単独では 1120 件のヒット件数，「溝口理一郎」単独では 1130 件のヒット件数であり，ほぼ同数であるから，「松尾豊」と AND をとったときの件数の違いは，氏名の共起関係の強さの違いを表していると考えることができる．すなわち，「松尾豊」と「石塚満」の方が，「松尾豊」と「溝口理一郎」よりも同一ページに出現する傾向が強い．したがって，関係が強いであろうことが推測される．実際，この例では，石塚満氏は松尾豊氏の学生時代の指導教官である．なお，本論文では，同一の Web ページに氏名が同時に現れることを，氏名が共起する，ということにする．

このように，本手法では基本的に，Web 文書における氏名の共起の強さによって関係の強さを推測する．氏名

*1 厳密には発表論文の著者と共著者で，聴講のみの参加者は含まない．聴講者を含む学会の参加者リストは学会主催者側は入手可能であるが，聴講者は参加することが公開されることを了承しているわけではないので，我々は Web 上に公開されている情報だけから氏名のリストを得るという方針を取っている．

*2 例えば，ReaD 研究開発支援総合ディレクトリ (http://read.jst.go.jp/) や J-STAGE の NII 学会発表データベース (http://www.jstage.jst.go.jp/) から入手可能である．

*3 2004 年 1 月 8 日時点での Google による検索結果．以下の例でも同様．Google では姓と名の間をつめて正確な氏名の検索が可能である．

が共起するページというのは，研究室のメンバーのページ，業績リストのページ，論文データベース，学会や研究会のプログラム，大学内の教官メンバーリストなどさまざまである．そして，このようなページが多くあるほど，両者が何らかの社会的関係にあり，またその関係が強い可能性が高い，というのが本研究の仮説である．この仮説は 5 章で検証する．

2·2 共起の強さを正確に知る

氏名の共起の強さを知るために，両者の名前の AND をとりヒット件数（共起頻度）を得ることは有用である．しかし，それを単純に関係の強さの推測値とするのは問題がある．

共起の強さを測るために，共起頻度以外にもさまざまな指標がある．集合の類似度，重なり具合を表す指標として，下記のようにさまざまなものが提案されている [Manning 02][Rasmussen 92]．ここでは，氏名「X」と氏名「Y」の単独でのヒット件数をそれぞれ $|X|$，$|Y|$，AND をとったとき，OR をとったときのヒット件数をそれぞれ $|X \cap Y|$，$|X \cup Y|$，Web ページ全体の数を N とする．

共起頻度

$$F(X,Y) = |X \cap Y| \tag{1}$$

相互情報量 $log\dfrac{N|X \cap Y|}{|X||Y|}$

ダイス係数 $\dfrac{2|X \cap Y|}{|X| + |Y|}$

Jaccard 係数 $\dfrac{|X \cap Y|}{|X \cup Y|}$

Simpson 係数 $\dfrac{|X \cap Y|}{\min(|X|, |Y|)}$

コサイン $\dfrac{|X \cap Y|}{\sqrt{|X||Y|}}$

共起頻度は，単独でのヒット件数が多い人ほど有利という問題がある．それに対して，他の係数は逆の欠点がある．仮に $|X|$ と $|Y|$ の差が大きい場合を考えよう．例えば，$|X| = 1000$, $|Y| = 30$, $|X \cap Y| = 30$ とすると，Jaccard 係数は 30/1000 と小さな値になる．$|Y|$ から見ると，すべてのページで $|X|$ と共起しているにも関わらず，値が小さい．逆にいうと，単独でのヒット件数が大きい有名な先生ほど，どの人とも関係が薄くなる傾向になる．

ただし，Simpson 係数は，分母に関して min をとっているので，この欠点がない．この係数は，ヒット件数の小さい方から見た距離感を表しており，例えば，研究室の学生と先生の場合にも，学生から見て先生と共起する割合が高ければエッジが張られることになるので，先生がたくさんのエッジを集めることになる．これは，研究室における協働関係に対して，我々の持っている印象と一致する．

しかし，Simpson 係数にも明らかな欠点がある．それは，単独でのヒット数が非常に少ない人には特に高い値が出やすいという点である．例えば，$|X| = 1$，$|Y| = 100$，$|X \cap Y| = 1$ の場合，Simpson 係数は 1 である．この欠点を解消するために，我々は次のような閾値つき Simpson 係数を用いることにした．

$$R(X,Y) = \begin{cases} \dfrac{|X \cap Y|}{min(|X|, |Y|)} & \text{if } |X| > k \text{ and } |Y| > k, \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

ただし，$R(X,Y)$ は，「X」と「Y」の関係の強さを表す関数であり，k は閾値である．JSAI2003 の場合，$k = 30$ とした．ヒット件数が低い人に対しても関係の強さをできるだけ的確に把握するには，例えば統計的な信頼度を推定する，m-estimate 法を用いるなどの方法が考えられる．しかし，ここでは簡単のため，閾値による足切りを行う上式を採用した．

$R(X,Y)$ は，ネットワークを構築する際に，閾値より高ければエッジを張り，そうでなければエッジを張らないという基準に用いる．また，エッジの長さとして用いることもできる．なお，この係数が適切であることは 5·1 節に述べる．

2·3 同姓同名の問題

単独でのヒット件数，例えば「松尾豊」のヒット件数を得る際，単純に

“松尾豊”

をクエリとして検索を行うと，容易に想像がつくように同姓同名の松尾豊氏が多数ヒットする．例えば，「松尾豊」の場合，ヒット件数 903 件に対し，本論文の著者である松尾豊氏に関するページは 256 件である．

このように，単独でのヒット件数 $|X|$ が，実際の値以上に大きくなってしまうと，式 (2) で定義される関係が，実際よりも小さな値になってしまう．そこで，氏名とともに所属情報を用い，より正確なヒット件数を得るという工夫を行う．例えば，「松尾豊」単独のヒット件数を得るために，

“松尾豊 産業技術総合研究所”

というクエリを用い検索する．JSAI2003 の場合には，学会のプログラムに記載されている所属情報を用いた．

また，著者によっては次のような問題があるため，それぞれに下記のように対応した．

複数の所属機関にまたがっている場合 複数の所属機関を OR でつないで検索を行う．“松尾豊 (産業技術総合研究所 OR 科学技術振興事業団)” という形でクエリを記述する．

所属機関名が複数存在する場合 略称，別名など．例えば，{ 東京大学，東大 }，{ 産業技術総合研究所，産総研，AIST} などは同義語である．このような代表的な機関の略称や別名については，同義語辞書を作り，同義語拡張を行った上で検索を行う．例えば，“

友部博教 (東京大学 OR 東大)” という形でクエリを記述する．

最近所属が変わった場合 過去の情報を含めて検索を行いたいので，過去の所属も OR でつないでクエリとする．JSAI2003 では，過去 5 年のプログラム上の所属の全てを用いる．例えば，“橋田浩一 (産業技術総合研究所 OR 電子技術総合研究所)” という形でクエリを記述する．

上記の 3 つに複数あてはまれば，複合的にクエリを拡張する．例えば,「松尾豊」であれば，実際には

“松尾豊 (産総研 OR 産業技術総合研究所 OR 東大 OR 東京大学)

で検索を行う.「松尾豊」単独で 903 件ヒットするのに対し，このクエリ拡張により 262 件にしぼられる．実際の正解のページ (本論文の共著者である「松尾豊」に関するページ) は 256 件であった．適合率は 86%，再現率は 93%となり，ほぼ目的とするページ群に近いものが得られていることがわかる．

もちろん，所属情報によるクエリ拡張を行っても，目的とする人物のページが 100%の適合率・再現率で検索できるわけではない．しかし，クエリ拡張を行わない場合に比べて，目的とする人物のページが格段に正確に得られる．

一方，氏名の共起 $|X \cap Y|$ に関しては，クエリに所属情報は用いない．例えば,「友部博教 AND 石塚満」の場合，人工知能の分野以外に，このペアでよく出現する同姓同名の人が存在する分野があれば，実際に求めたいもの以上にヒット件数が多くなり，問題となる．しかし，氏名単独での同姓同名に比べて，氏名のペアが同姓同名である確率は低い．また，人工知能学会では，我々の知る限り同姓同名はいないため，ここでは大きな問題がないと考え，氏名を AND でつないだものをそのまま用いている．Google は 10 個以上のクエリーを受け付けないことも所属情報を用いない 1 つの理由である．なお，Simpson 係数（式 (2)）において，$|X|$，$|Y|$ は所属情報を用い，$|X \cap Y|$ は所属情報を用いないので，結果的に Simpson 係数が 1 を越えることもある．

3. エッジラベルの抽出

前章では，検索エンジンを用いて，Web ページにおける氏名の共起の強さから 2 人の関係の強さを推測する方法を示した．では，2 人がどのような社会的関係にあるかを知ることはできないだろうか？本章では，氏名が共起したページ，つまり検索にヒットしたページの特徴を用いて，関係の種類を判別する手法について述べる．

3·1 関係を判別する方法

社会的関係の種類として，本論文では研究分野に特有の次のようなクラスを定める．これらが，エッジのラベルの種類となる．

表 2 語群

語群	語
A	出版， 論文， 発表，活動，テーマ， 賞， 著者
B	メンバー，研究室，研究所，研究機関，チーム
C	プロジェクト, 委員会
D	ワークショップ, 会議, セミナー, ミーティング, スポンサー, シンポジウム
E	学会, 団体, プログラム, 国立, ジャーナル, セッション
F	教授, 専攻, 大学院生, 講義

共著関係 共著の論文がある関係.

同研究室関係 同じ研究室や研究所のメンバーなど所属が同じである (あった) 関係.

同プロジェクト関係 同じプロジェクトや委員会など，組織をまたがる同グループに所属している (いた) 関係.

同発表関係 同じ研究会で発表する (した) 関係.

以後，誤解のない範囲で,「共著」「研究室」「プロジェクト」「発表」と略記する．ひとつのエッジは複数のラベルを持つことができる．今回は，研究者を対象としているのでこのような関係を定義したが，一般的には対象とする領域ごとに定義する必要がある．

さて，このような関係を抽出するために，まず検索エンジンに「X and Y」をクエリとして入力し，上位 5 ページを取得する[*4]．次に，それぞれのページから表 1 にある属性の値を抽出する．属性 NumCo，FreqX，FreqY は，氏名 X，Y のページ内での出現に関する属性である．一方，GroTitle 属性，GroFFive 属性は，そのページが何に関するページであるかを判断するためのものであり，別に定義した語群（表 2）を用い，語群 A がタイトルに出現するかどうか（GroTitle(A) 属性），語群 B が最初の 5 行に出現するか（GroTitle(B) 属性）などを表す．各語群は手動で設定する方法もあるが，できるだけ自動化するために，あらかじめ正解クラスの付与されたページを用い，各クラスごとに TF·IDF 値の上位語[*5]を語群としている．

例を用いて説明すると,「友部博教 AND 石塚満」で検索したあるページ[*6]から表 1 の属性を抽出すると，

(more_than_one, yes, yes, more_than_one, more_than_one, no, no, no, no, no, no, yes, no, no, no, yes, no)

となる．そして，この属性から共著・研究室・プロジェクト・発表という 4 つのクラスに属するかどうか，このページの場合には

(Yes, No, No, Yes)

を得るという問題になる．さらに検索にヒットした他のページからも関係を求め，最終的に 2 人の関係のエッジラベルとして

*4 Google の上位候補は PageRank が高い Authority ページであり，またなるべく重複が避けられるように工夫されているので，そのまま上位から優先的に用いる．

*5 TF·IDF 値が 15 を越える語．この値は経験的に定めた．

*6 http://www-kasm.nii.ac.jp/jsai2003/programs/person-182.html

表 1 「X AND Y」でヒットした Web ページから抽出する属性

属性名	説明	値
NumCo	二人の氏名の共起回数	zero, one, more_than_one
Rel	式 (2) で表される Simpson 係数が閾値以上か	yes, no
FreqX	X の出現回数	zero, one, more_than_one
FreqY	Y の出現回数	zero, one, more_than_one
GroTitle	タイトルに語群（A-F）が出現するか	（語群 A-F に対してそれぞれ）yes, no
GroFFive	最初の 5 行に語群（A-F）が出現するか	（語群 A-F に対してそれぞれ）yes, no

表 3 獲得した全判別ルール

クラス	判別ルール
共著	NumCo = more_than_one
研究室	(NumCo = more_than_one & GroFFive(F) = no) or (Rel = yes & GroTitle(E) = no & GroFFive(C) = no) or (GroTitle(A) = yes & GroFFive(C) = no & GroFFive(F) = yes) or (GroTitle(E) = no & GorFFive(B) = yes & GroFFive(C) = no) or (GroTitle(E) = no & GroFFive(B) = yes & GroFFive(E) = yes) or (GroTitle(E) = no & GroFFive(B) = yes & GroFFive(F) = yes)
プロジェクト	(FreqX = one & GroTitle(B) = yes) or (GroTitle(C) = yes) or (GroTitle(C) = no & GroFFive(C) = yes & GroFFIve(D) = no & GroFFive(E) = no) or (Rel = no & FreqX = one & GroTitle(B) = yes)
発表	(FreqY = more_than_one & GroTitle(D) = yes) or (GroTitle(F) = yes & GroFFive(D) = yes) or (NumCo = zero & GroTitle(F) = no & GroFFive(B) = no & GroFFive(E) = no & GroFFIve(F) = yes) or (NumCO = zero & GroTitle(C) = no & GroFFive(D) = yes & GroFFIve(E) = no) or (NumCO = zero & GroTitle(C) = no & GroTTl(D) = no & GroFFive(D) = yes)

(Yes, Yes, No, Yes)

つまり，共著かつ研究室かつ発表関係であると求めたい．

したがって，ページの属性から自動的にクラスを判別できればよい．これは，属性からクラスを予測するルールを学習する問題となる．本研究では，C4.5[Quinlan 93] を用いて判別ルールを生成する．ランダムに抽出した 275 ページを人手で正解クラスを付与し，これを訓練例として用いた．なお，SVM など他の学習アルゴリズムを用いることもできるが，結果の解釈の容易性を確保するため，C4.5 を用いている．獲得したルールを表 3 に示す．

まず，共著関係のルールでは，氏名が同行内に出現することが 2 回以上あれば（NumCo=more_than_one）共著と判断する，という非常に簡単なものである．研究室関係を判断するルールでは，例えば 1 つ目のルールは，名前が 2 回以上共起している（NumCo=more_than_one）のに学科や講義のページではなければ（GroFFive(F)=no）研究室関係である，というルールである．2 つ目のルールは，関係が強い（Rel=yes）のにプロジェクトでも研究会でもなければ（GroFFive(E)=no & GroFFive(C)=no）研究室関係であるというルールである．このように，各クラスに対していくつかの判別ルールが得られる．

3･2 判別ルールの精度

表 4 に，275 の訓練例を 5 群に分け，クロスバリデーションを行った平均エラー率を示す．研究室クラスに対するエラー率が高いが，他のクラスでは 10%程度もしくはそれ以下のエラー率である．また，実際に得られたラベルの適合率，再現率を知るために，ルールを生成する際に用いた 275 の訓練例とは別にランダムに 200 個のエッジを選び，そのラベルを人手で判定し，適合率，再現率を求めた．なお，この評価はネットワークを生成した後で行ったものであり，評価のための 200 例はルール生成には利用していない．

共著クラスは，非常に簡単なルールであるが，適合率，再現率ともに高い．一方，研究室クラスの適合率は低いが，これは個人の業績のページで，人間が見れば共著関係か研究室関係か（その両方か）は区別できても，ここで用いている属性だけでは判別しづらいためであると考えられる．プロジェクトクラス，発表クラスでは，クロスバリデーションのエラー率と比較し，200 例に対する適合率，再現率から算出したエラー率は高い．これは，プロジェクトや発表を表すページが比較的多様であり，結果的に 275 例では訓練データが偏り，十分に学習できていないためと考えられる．

表 4 ラベルのエラー率，適合率と再現率

クラス	エラー率*	適合率	再現率	エラー率**
共著	4.1%	91.8% (90/98)	97.8% (90/92)	5.0%
研究室	25.7%	70.9% (73/103)	86.9% (73/84)	20.5%
プロジェクト	5.8%	74.4% (67/90)	91.8% (67/73)	14.5%
発表	11.2%	89.7% (87/97)	67.4% (87/129)	26.0%

*: 225 訓練例のクロスバリデーションによるエラー率
**: 別の 200 例に与えたときのエラー率

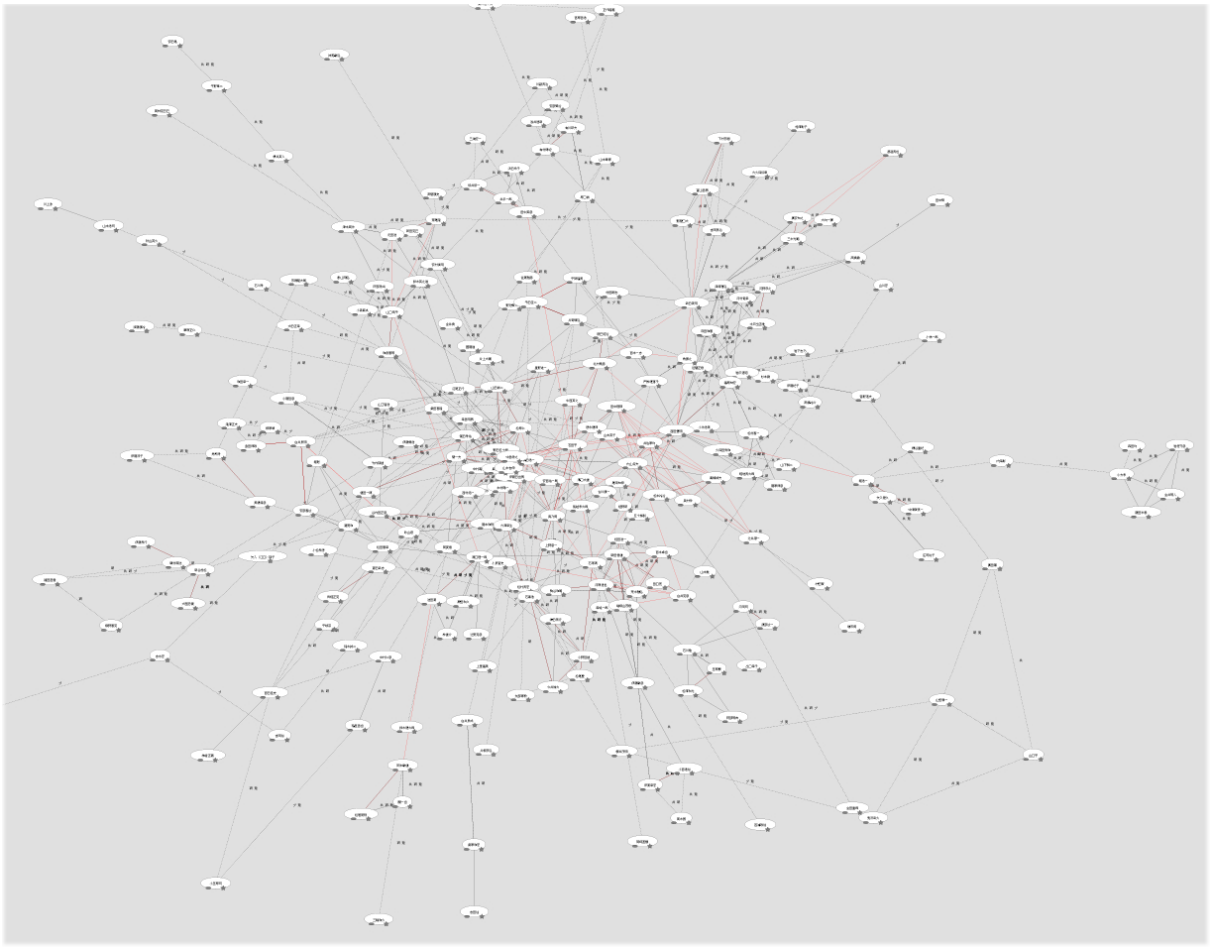

図 1 JSAI2003 で表示した人間関係ネットワーク

4. 人間関係ネットワークの表示

JSAI2003 では，2 章および 3 章の手法により得たネットワークを，会場内に設置された KIOSK 端末および Web 上で表示するサービスを行った[*7]．図 2 はその様子である．表示したネットワークを図 1 に示す．ノード数 266，エッジ数 690[*8]のネットワークであり，一部を拡大したものが図 3 である．なお，2003 年および過去 4 年の人工知能学会で発表した著者，共著者は合計 1560 人であり，そのなかで単独でのヒット件数が 30 件に満たない，もしくは他のノードとの Simpson 係数が閾値（0.5）以下で孤立ノードとなってしまうノードは除外し，結果的に 266 ノードとなった．

ネットワークは，SVG[*9]で出力され，SVG viewer により閲覧することができる．Javascript が埋め込まれているので，ノードをドラッグしてつながり具合を確かめることができる．各ノードには丸印と星印のアイコンがあり，それぞれスケジューリング支援システム，CoBIT による位置情報表示システムと連携している．（各システムの概要は [西村 04] 参照.）エッジは，Simpson 係数 $R(X,Y)$ が閾値（0.7）を越えるノードペア X，Y に対して実線で表示している．エッジラベルとして，“共”（共著），“研”（研究室），“プ”（プロジェクト），“発”（発表）がそれぞれ 243 本，243 本，92 本，192 本のエッジに付与されており，それらをクリックすると，その判断の根拠

*7 http://www.carc.aist.go.jp/~y.matsuo/humannet/ からネットワークにアクセスできる．
*8 実線エッジ 284，破線エッジ 262，赤エッジ 144．区別については後述．
*9 SVG は，W3C によって作成された規格であり，ベクトル表現による XML 形式のグラフィック記述言語である．

図 2 JSAI2003 会場における人間関係ネットワークの表示

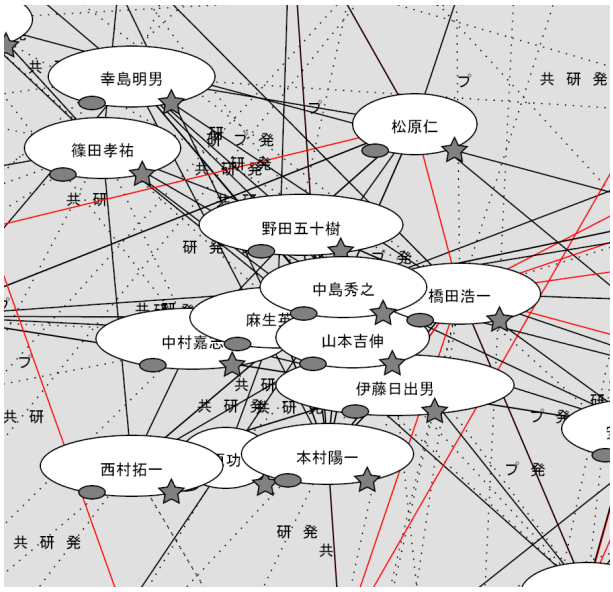

図 3 JSAI で表示した人間関係ネットワーク（拡大図）

となったページへジャンプする．初期配置では，エッジの長さが $R(X,Y)$（の逆数）をできるだけ反映するような配置となっている[*10]．

また，$R(X,Y)$ による黒い実線のエッジの他にも，ネットワーク表示によるコミュニケーションの促進となるように，次のような 2 種類のエッジを加えて表示した．

赤エッジ 共起頻度（式(1)）が閾値（100）以上のものについて赤色のエッジを表示している．ヒット件数の多い有名な人のペアが多く含まれるが，このようなエッジはコミュニティの骨格を表すために有用なものである．（なお，黒い実線と両方ある場合にはこちらを優先する．）

破線エッジ このネットワークは，関係性を図示してコミュニケーションを促進する目的があるので，各ノードに対してエッジが 3 本以下の場合，閾値をさらに下げて（0.5）破線でエッジを表示する．

JSAI2003 の会場で運用を行うことによって，旧姓の併用の問題，外国人名の問題など，いくつか問題点も明らかになったが，人間関係ネットワークを表示するページへのアクセスも多く，分かりやすく面白いシステムであった，研究者の全体的な関係を理解するのに役立ったなどの声も聞かれた．

5. 人間関係ネットワークの分析と評価

5·1 共起の指標の評価

2 章では氏名間の関係性の強さを，3 章では関係の種類を取得する手法について述べた．では，これらの間にはどのような関係があるだろうか？

ここでは，ラベルの抽出精度が最も高い共著関係を取り上げる．図 4 は，訓練例として用いた 275 例について，横軸に $R(X,Y)$ を，縦軸に共著関係である確率（前後 20 点の移動平均）をとったグラフである．$R(X,Y)$ が 0.2[*11] を越えると，ほぼ確実に共著の関係がある．一方，それ以下の場合，共著関係がある確率は急激に下がる．

一方，図 5 は，横軸に氏名の共起頻度 $F(X,Y)$ を，縦軸に共著関係である確率をとったものである．共起頻度が 50 を越えた領域でも，共著関係であるかどうかに揺れがある．これは，2·2 節で指摘したように，ヒット件数が多い人は一般に共起頻度は高くなる傾向があり，共著関係がないにも関わらず高い共起頻度となるためである．図 6 は共著関係と Jaccard 係数の関係を表している．係数の値が高い領域では安定して共著関係があるが，値が低い領域（0.01 近辺）でも，共著関係のある確率が突出している．これは，ヒット件数 $|X|$ と $|Y|$ に大きな差がある場合，関係が強くても係数の値が大きな値をとらないという，2·2 節で指摘した問題のためである．また，閾値なしの Simpson 係数（図 7）は逆に，係数の値が 0.2 を越えたあたりでも共著関係の有無が安定していない．2·2 節で述べたように，ヒット件数の非常に低い人に対して，係数値が大きくなる問題があるためである．

これらの図を見比べると，閾値つきの Simpson 係数（式(2)）は，係数が低い領域でも高い領域でも，共著関係が存在するかに関する安定した指標となっている．

5·2 アンケートによる評価

JSAI2003 の後，我々は人間関係ネットワークに関するアンケート調査を行った．調査対象者は，JSAI2003 に

*10 Graphviz(`http://www.research.att.com/sw/tools/graphviz/`) を使い，ばねモデルによる初期配置を求めている．

*11 なお，このグラフは係数の最大値が 1 となるように決めている．Simpson 係数は 2·3 節に述べた理由により 1 を越えることがある（この場合，最大 4.45 であった）ため，全データを [0,1] の間におさめるためにもとの Simpson 係数を 4.45 で割っている．したがって，ネットワークの図示に用いた閾値である 0.5，0.7 という値は，ここでは 0.11，0.16 にあたる．以下のグラフも，比較のため，最大値が 1 となるように正規化している．

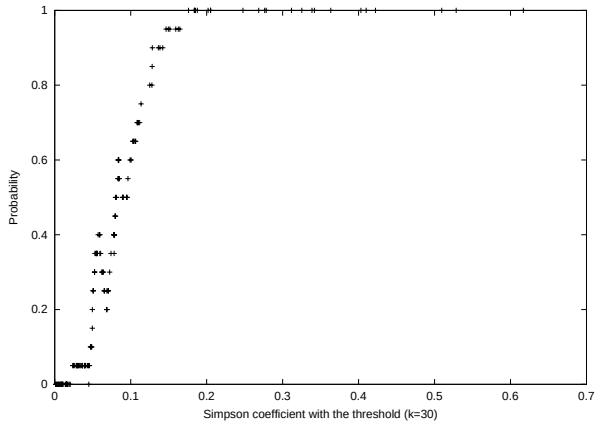

図 4 共著ラベルの出現率と閾値付き Simpson 係数

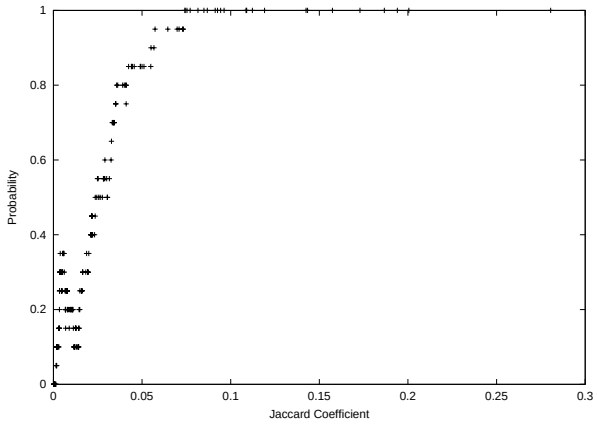

図 6 共著ラベルの出現率と Jaccard 係数

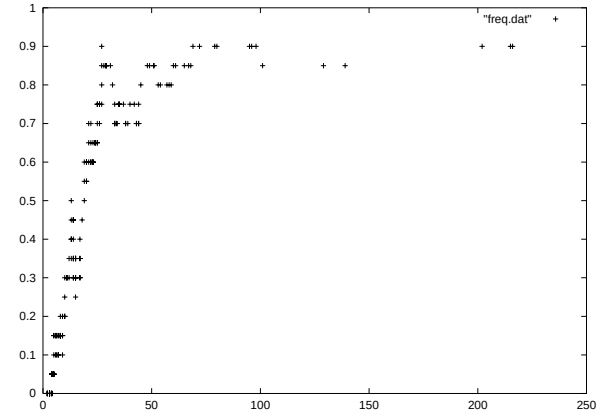

図 5 共著ラベルの出現率と共起頻度

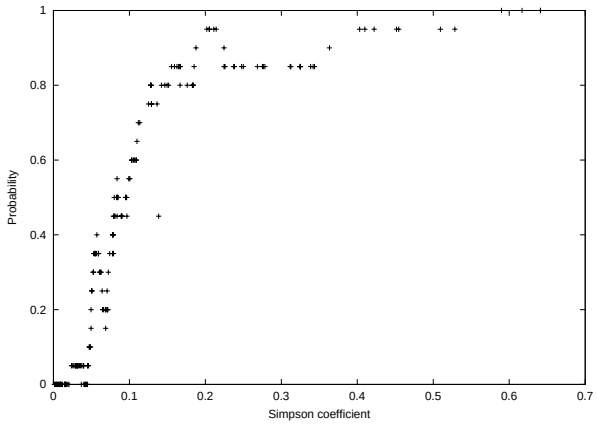

図 7 共著ラベルの出現率と Simpson 係数（閾値なし）

参加登録した人の中から選んだ 141 人とした[*12]．CGI によるアンケートシステムを作成し，アンケートの協力をお願いするメールを対象者に送付した．82 名から回答を得，回収率は 58%であった．

アンケートでは，各被験者に対して，スケジューリング支援システムで know リンクを張った／張られた人から 10 人，さらに我々のシステムにおいて共起の閾値つき Simpson 係数 $R(X,Y)$ に応じたルーレット選択で 10 人を抽出し，一人あたり各 20 人の相手との関係を質問した．know リンクとは，ユーザが明示的に「この人を知っている」と登録したリンク関係を指す．質問は，一人の相手あたり各 15 問であり，「共著の論文がある（既に公になっているものに限る）」，「同じ研究室や部署など 30 人規模の組織に同時期に所属している，またはしていた」，「同じプロジェクトや委員会に所属している，またはしていた」，「JSAI2003 以外の研究会や国際会議で会ったことがある」などの項目を含む．それぞれ，共著，研究室，プロジェクト，発表の関係の有無を問う意図で設定した質問項目である．

表 5 に，JSAI2003 で表示したネットワークのエッジラベルに対して，アンケートから得た回答を正解とした場合の適合率および再現率を示す．また，表 6 は，抽出した全関係に対する適合率および再現率であり，ネットワーク中にエッジラベルとして表示していないものも含む．言いかえれば，関係の強さ $R(X,Y)$ が閾値以下のノードペアに対するラベルも含んでいる．

表 4 と比較して，表 5，表 6 は，適合率，再現率ともに低い値になっている．この理由として考えられるのは，

- 回答者が共著やプロジェクトの関係を忘れている，記述もれしているなどの可能性がある．例えば，表 5 の共著で，システムの出力が誤っていたとされた 10 件 (= 91 件 – 81 件) 中，6 件は実際には共著の関係があった．特に発表関係は，はっきりと覚えていない場合も多いと考えられる．
- プロジェクトの定義としてより広いものを想定しており，再現率が低くなっている．

など，アンケートの回答に関する問題である．しかし，最も大きな原因，特に再現率が低いことに対する原因とし

*12 JSAI2003 にスケジューリングシステムにメールアドレスを登録した 231 人のうち，スケジューリングシステムにおける know リンクの数と本システムにおけるエッジの数の和が 10 人に達しない 90 人を除いた，141 人全員を対象者とした．アンケート送付は 2003 年 12 月 4 日であり，その後約 2 週間で回答を締め切った．

表 5 JSAI2003 ネットワークにおけるエッジラベルのアンケートによる評価

クラス	適合率	再現率
共著	89.0% (81/91)	32.1% (81/252)
研究室	78.3% (72/92)	18.7% (72/385)
プロジェクト	50.0% (9/18)	3.0% (9/300)
発表	79.5% (35/44)	6.5% (35/538)

表 6 抽出された全エッジラベルのアンケートによる評価

クラス	適合率	再現率
共著	78.5% (135/172)	53.6% (135/252)
研究室	55.6% (109/198)	28.3% (109/385)
プロジェクト	20.3% (60/296)	20.0% (60/300)
発表	39.9% (222/556)	41.3% (222/538)

て考えられるのは，次のような点である．

- 我々の手法は，すべての Web ページを網羅的に分析しているわけではない．検索でヒットした上位 5 ページのテキストを分析したものであるため，関係を取り逃す場合もある．
- すべての情報が Web 上にあるわけではない．プロジェクトに関しては，Web 上に情報がないものも多い．また，研究室には過去のメンバーリストを載せておらず，現在のものに書き変わっている場合もある．

そもそも，Web 上にない情報から関係を把握することはできないので，本論文のアプローチが再現率に対して限界があることは明らかである．しかし，関係の強さと併せて用いることで，表 5 で示したように 80%程度の適合率で共著，研究室，発表などの関係を抽出できるということは，学会におけるコミュニケーション支援という目的には有用であろう．また，本論文で用いた方法はシンプルなものであり，属性の取り方や学習の方法を工夫することで，さらに高い精度を得ることも可能であると考えられる．

6. 議　　論

6·1 Web 上の情報の可能性と限界

これまで，論文データベースのデータを用い，論文の共著者関係や引用関係，共引用関係からネットワークを構築する研究は長年に渡って行われてきた [Yoshikane 04, Garfield 64]．しかし，本手法は，共著者という関係も含んだ上で，研究室やプロジェクト，発表などの関係も抽出することができる．野村らは，研究者の Web サイトの引用解析を行い，Web と文献引用の解析を比較した結果，Web の共引用によるクラスタが文献の共引用によるクラスタを含む傾向があることを明らかにしている [野村 04]．Web 上の情報は，論文データベースと比べ，整形されていない雑多で多様な情報であるがゆえに，より複雑な処理を必要とするが，多様な関係を抽出できる可能性がある．

本手法は，あくまでも Web 上に情報がなければその関係を取得することは不可能である．しかし，今後ますます多くの情報が Web 上に公開される流れにあるとすると，本手法の適用範囲は広がるだろう．

抽出した人間関係ネットワークは，そのネットワークの内部にいる人からみれば，当たり前の関係であるかもしれないし，漏れも多いかもしれない．しかし，当分野の学生や若手研究者，また他分野の研究者にとって，領域を俯瞰し自分の立場や周りの関係を知ることは有用であろう．特に分野間の融合や分野横断的な研究が求められる領域では特に重要な技術である．また，情報支援システムがこのような人間関係ネットワークに関する情報を背景知識として持てば，個人のネットワーク上の立場に応じたさまざまな情報支援が可能となるだろう．

6·2 プライバシーについて

さて，本手法では，氏名と所属情報だけをもとにネットワークを構築している．プライバシーには十分配慮を行っているが，個人情報を扱う際にはさまざまな点に注意する必要がある．まず，一般的な意味での「人間関係」はプライバシーに該当する．他者が知り得ない情報を許可なく公開するのは違法性が高い．しかし，本研究で扱っているのは，一般的な意味での人間関係ではなく，研究者の協働関係であり，Web 上に公開されている他者が知り得る情報をもとにしている．ただし，公開されている情報であるからといって，自由に用いてよいものではない．情報を提供している人が想定する範囲を越えて情報を流通させることは問題がある．本研究の場合，論文を研究会やジャーナルに投稿することや，研究室のページにメンバーリストを載せることが，関連する研究者に共著者や研究室メンバーとの関係を知らしめることになるのは，当の研究者の想定範囲内であると考えられる．したがって，本研究で行っているネットワークの図示は，プライバシーの侵害にあたるとは考えていない．特に，本手法は，メールのやりとりからネットワークを構築する，アンケート調査によりネットワークを構築し公開するなどと比べると，もとの情報がオープンであり情報提供者が意図している範囲内であると考えられることからプライバシーの問題は少ないが，運用上はこの点について今後も配慮を行っていくつもりである．

6·3 関 連 研 究

これまでに挙げた以外の関連研究として次のものが挙げられる．Referral Web[Kautz 97] は，自分から対象人物へのつながりを Web 上の情報から順次発見していくものである．重要な情報は Web ではなく人が持っているので，人のつながりを見つけることが重要であるなど，本研究と問題意識は近い．しかし，本研究との違いは，名前のリストを最初に与えるわけではなく，ひとつの名前と共起する名前を抽出し，さらにその名前から次の名前を抽出するというように反復的にネットワークを広げて

いくため，取り出したいコミュニティの人間関係が出るわけではない．また，あるノードから距離 3 のノードを収集するのに 24 時間程度かかる．さらに，エッジのラベルを考慮していない，評価を行っていない，本研究では問題があると示した Jaccard 係数を用いている，などの点で本研究と異なる．

原田らは，ある単語で検索した Web ページ集合 (最大 1000 件) から固有表現抽出により名前を抽出する [原田 03]．そして，独自に定義した共起度を用い，共起関係から人物の関係を表すネットワークを抽出する方法を提案している．しかし，ある「分野」の人物の関係ネットワークであり，与えられた名前リストのネットワークではない点，エッジのラベルを考慮していない点，評価を行っていない点で本研究と異なる．

一方，村田らは検索エンジンで検索された件数を用いて Web ページ間の関係を発見する手法を提案している [村田 01] が，人間を扱っているという点で大きく異なる．また，Web 上の情報だけでなく，社内のメールのやり取りを用いて人間関係を抽出する研究も行われている [Tyler 03, Guimera 02]．しかし，この手法は情報源にプライバシーの問題があり，それを使った情報支援をスケールさせていくことは難しい．

7. あ と が き

本論文では，Web 上の情報からの人間関係ネットワークの抽出と題して，学会というコミュニティにおける協働関係を抽出する手法を述べた．非常にシンプルなアルゴリズムでありながら，効果的に関係を抽出することができる．

このような人間関係に関する情報は，リアルワールドでの情報支援を考えると非常に重要である．実際,「誰といるか」はその人の文脈を決定する大きな要素であると考えられるから，ユビキタス環境における情報支援のひとつの背景知識として使うと，さまざまな面白い試みが可能であろう．

本手法は，ある研究分野における研究活動を分析する，促進するといった点でも，重要な基礎データになると考えられる．また，近年では，特に Semantic Web のトラストレイヤーに関わる話題で，人間ネットワークが取り上げられることが多い．本論文で提案する自動抽出法は，ネットワークを網羅的にかつ一様な基準で抽出できるという点で，ひとつの有効なアプローチであると考えている．

謝　　辞

JSAI2003 におけるイベント支援において，産業技術総合研究所 西村 拓一氏，国立情報学研究所 武田英明氏，濱崎雅弘氏，大向一輝氏，東京大学 森純一郎氏にさまざまな形で御協力を頂きました．また，本論文をまとめるにあたり，GBRC 社会ネットワーク研究所所長 安田雪氏には的確なアドバイスを頂きました．ありがとうございました．

◇ 参　考　文　献 ◇

[Brin 98] Brin, S. and Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine, in *Proc. 7th WWW Conf.* (1998)

[FOAF] FOAF, : FOAF: the 'friend of a friend' vocabulary, http://xmlns.com/foaf/0.1/

[Garfield 64] Garfield, E., Sher, I., and Torpie, R.: The use of citation data in writing the history of science, Technical report, Philadelphia Institute of Scientific Information (1964)

[Guimera 02] Guimera, R., Danon, L., Diaz-Guilera, A., Giralt, F., and Arenas, A.: Self-similar Community Structure in Organizations, ArXiv:cond-mat/0211498 v1 (2002)

[Kautz 97] Kautz, H., Selman, B., and Shah, M.: The Hidden Web, *AI magazine*, Vol. 18, No. 2, pp. 27–35 (1997)

[Kleinberg 98] Kleinberg, J. M.: Authoritative Sources in a Hyperlinked Environment, *Proc. ACM-SIAM Symposium on Discrete Algorithms*, pp. 668–677 (1998)

[Kumar 99] Kumar, S. R., Raghavan, P., Rajagopalan, S., and Tokins, A.: Trawling the web for emerging cyber communities, in *Proc. 8th WWW Conf.* (1999)

[Manning 02] Manning, C. D. and Schütze, H.: *Foundations of statistical natural language processing*, The MIT Press, London (2002)

[村田 01] 村田 剛志：参照の共起性に基づく Web コミュニティの発見, 人工知能学会誌, Vol. 16, No. 3, pp. 316–323 (2001)

[中島 01] 中島, 橋田, 森, 伊東, 本村, 車谷, 山本, 和泉, 野田：情報インフラに基づくグラウンディングとその応用 – サイバーアシストプロジェクトの概要 –, コンピュータソフトウェア, Vol. 18, No. 4, pp. 48–56 (2001)

[Quinlan 93] Quinlan, J. R.: *C4.5: Programs for Machine Learning*, Morgan Kaufmann, California (1993)

[Rafiei 00] Rafiei, D. and Mendelzon, A.: What is this Page Known for? Computing Web Page Reputations, in *Proc. 9th WWW Conf.* (2000)

[Rasmussen 92] Rasmussen, E.: Clustering Algorithms, *Information Retrieval: Data Structures & Algorithms. William B. Frakes and Ricardo Baeza-Yates (Eds.), Prentice Hall* (1992)

[Tyler 03] Tyler, J., Wikinson, D., and Huberman, B.: *Email as spectroscopy: automated discovery of community structure within organizations*, pp. 81–96, Kluwer, B.V. (2003)

[Yoshikane 04] Yoshikane, F. and Kageura, K.: Comparative analysis of coauthorship networks of different domains: the growth and change of networks, *Scientometrics*, Vol. 60, No. 3, pp. 435–446 (2004)

[安田 97] 安田 雪：社会ネットワーク分析 –何が行為を決定するか–, 新曜社 (1997)

[原田 03] 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, 情報処理学会研究報告, 第 DBS-130/FI-71 巻 (2003)

[西村 04] 西村 拓一, 濱崎 雅弘, 松尾 豊, 大向 一輝, 友部 博教, 武田 英明：2003 年度人工知能学会全国大会支援統合システム, 人工知能学会誌, Vol. 19, No. 1, pp. 43–51 (2004)

[野村 04] 野村 早恵子, 三木 武, 石田 亨：コミュニティマイニングにおける Web 引用解析と文献引用解析の比較, 電子情報通信学会論文誌 D-I, Vol. J87-D-I, No. 3, pp. 382–389 (2004)

[濱崎 02] 濱崎 雅弘, 武田 英明, 松塚 健, 谷口 雄一郎, 河野 恭之, 木戸出 正継：Bookmark からの共通話題ネットワークの発見手法の提案とその評価, 人工知能学会論文誌, Vol. 17, No. 3, pp. 276–284 (2002)

〔担当委員：庄司裕子〕

2004 年 4 月 14 日　受理

著　者　紹　介

松尾　　豊（正会員）

1997 年東京大学工学部電子情報工学科卒業．2002 年同大学院博士課程修了．博士（工学）．同年より，産業技術総合研究所サイバーアシスト研究センター勤務．2004 年 7 月より産業技術総合研究所情報技術研究部門．2002 年度人工知能学会論文賞受賞．推論，キーワード抽出，Web マイニング等に興味がある．受け手にとって価値の高い情報の提示を目指している．情報処理学会，AAAI の各会員．

友部　博教（正会員）

1999 年東京大学工学部電子情報工学科卒業．2004 年同大学院情報理工学系研究科博士課程修了．博士 (情報理工学)．現在，名古屋大学 21 世紀 COE プログラム「社会情報基盤のための音声・映像の知的統合」ポスドク．マルチメディアコンテンツからの知識発見等の研究に従事．情報処理学会会員

橋田　浩一（正会員）

1981 年東京大学理学部情報科学科卒業．1986 年同大学院理学系研究科博士課程修了．理学博士．1986 年電子技術総合研究所入所．1988 年から 1992 年まで (財) 新世代コンピュータ技術開発機構に出向．2001 年から産業技術総合研究所サイバーアシスト研究センター副研究センター長，ついで研究センター長．2004 年 7 月より産業技術総合研究所情報技術研究部門副部門長．専門は自然言語処理，人工知能，認知科学．現在の研究テーマはセマンティックコンピューティングおよびそれに基づく知の社会的共創など．

中島　秀之（正会員）

1983 年，東京大学大学院情報工学専門課程修了（工学博士）．人工知能を状況依存性の観点から研究．最近は情報処理の社会システムへの応用に興味を持っている．現在，公立はこだて未来大学学長．産業技術総合研究所情報技術研究部門研究顧問．認知科学会会長，ソフトウェア科学会元理事，人工知能学会元理事，情報処理学会元理事．マルチエージェントシステム国際財団元理事．主要編著書：AI 事典第 2 版（共立出版），知的エージェントのための集合と論理（共立出版），思考（岩波講座認知科学 8），記号の世界（岩波書店），Prolog（産業図書）．

石塚　　満（正会員）

1971 年東京大学工学部電子卒，1976 年同大学院博士修了．同年 NTT 入社，横須賀研究所勤務．1978 年東京大学生産技術研究所・助教授，(1980-81 年 Purdue 大学客員準教授)，1992 年東京大学工学部電子情報・教授，2001 年情報理工学系研究科電子情報学専攻，現在に至る．研究分野は人工知能，インターネット/WWW インテリジェンス，生命的エージェントによるマルチモーダルシステム．IEEE, AAAI, 情報処理学会，電子情報通信学会，映像情報メディア学会，画像電子学会，等の会員．