論　文

Web上の情報を用いた弱い社会的関係のネットワーク抽出手法

金　英子[†a)]　松尾　豊[††]　石塚　満[†]

Extracting Social Networks with Weak Relationships from the World Wide Web

Yingzi JIN[†a)], Yutaka MATSUO[††], and Mitsuru ISHIZUKA[†]

あらまし　近年，Web上から人間関係ネットワークを抽出し，可視化，情報共有，分析等に利用する研究が行われている．従来の研究では，Webから人間関係の強さをどのように計量化するかについて検討されてきたが，いずれの研究でもネットワーク全体に一貫した基準で関係の有無を判断しエッジを張ることで，ネットワークを構築していた．その結果，対象とするコミュニティによってはネットワークの一部分にエッジが集中し，その他のエッジを適切に抽出することができないという問題があった．本研究では，ネットワーク全体からみて弱い社会的関係であっても，その人にとって相対的に強い関係の人々を見つける新たな人間関係抽出の方法を提案する．4つのパラメータを調整することで，適切な抽出が可能であることを示す．本研究の手法は，横浜トリエンナーレのアーティストのネットワークを抽出するために用いられ，ユーザをナビゲートするWebサイトとして運用された．

キーワード　Webマイニング，社会ネットワーク，関係抽出，弱い関係

1. まえがき

社会ネットワーク分析は，人と人との関係性に注目し，そのつながりを測定してネットワーク化することにより，社会構造や現象を説明する社会学におけるひとつの方法論であり[4], [13], [15]，1940年代から研究されている．近年，人のつながりや社会性が情報技術分野でも注目されており，社会ネットワーク分析と情報技術をつなぐさまざまな研究が行われている（例えば[6], [12]）．

人のつながりをネットワークとして抽出する，すなわちネットワーク化するには，だれがだれとどのような関係にあるかというデータを取得する必要がある．社会学では従来，インタビュー調査や長時間の観察によってこういったデータを集めていた．人のネットワークを抽出するために用いられる代表的な方法は，ネットワーク・クエスチョンによる調査である[15]．例えば，アメリカのGSS（General Social Survey）調査では，「あなたが過去半年のあいだに，あなたにとって重要なことを話しあった人々は誰でしたか？」というネットワーク質問を行う．このような質問により，個人のもつネットワークや関係性の分析を行うことができる．しかし，多くの人に対して定期的にこのようなアンケートを行うことは難しい．

一方，近年活発に研究されているのは，電子化された情報をもとに社会ネットワークを抽出する方法である．850億ものページ(注1)が存在するWebは，ある種の人間社会を反映していると言っても過言でない．日常話題になっているニュースや学会・展示会などのイベント，個人のホームページやBlogなど，Web上には社会的な活動に関する多岐にわたる情報が存在する．Web上の情報は，社会学者の注目も集めており[14]，Webから社会ネットワークを抽出し分析する研究が盛んに行われている．

Webから社会ネットワークを抽出する早い時期の研究としては，Kautzらの Referral Webがある．このシステムでは，ユーザはWeb上の名前の共起関係から

† 東京大学大学院　情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, 113-8656 Japan
†† 東京大学大学院　工学系研究科
Graduate School of Engineering, The University of Tokyo, Tokyo, 113-8656 Japan
a) E-mail: eiko-kin@mi.ci.i.u-tokyo.ac.jp

(注1)：InternetArchiveのWayback Machineが1996年から2007年初頭までに収集したWebページ数（http://www.archive.org/web/web.php, About the Wayback Machine）．

研究者のつながり（例えば Henry Kautz から Marvin Minsky へのパス）を発見することができる [5]．また，2005 年から 2006 年にかけて研究開発された Mika らの Flink というシステムは，Semantic Web に関係する研究者のネットワークを抽出する [10]．同様に，松尾らは Web 上の名前の共起関係，及び共起のコンテキストから研究者の関係を把握する Polyphonet というシステムを開発し，人工知能学会などの学会においてコミュニケーションの支援を行っている [8], [9]．他にも，ユーザが入力したトピックに関連する人物の関係を Web から抽出する原田らの研究 [3]，E-mail と Web 上の情報の両方を用いて人物の関係を取得する Bekkerman らの研究 [1]，Web 上の引用・共引用の情報を用いて関係を抽出する Miki らの研究 [11] もある．本研究は，このような一連の研究のひとつとして位置づけられる．特に，現代美術，パフォーマンス，建築などに関するアーティストのネットワークを抽出することが目的であり，そのためにネットワークの構成法に関して従来考慮されていなかった点を扱っている．

Web 上から人間関係を抽出する既存の研究（具体的には [3], [5], [8]〜[10]）では，人の関係の強さを把握する際に検索エンジンのヒット件数を用いる．例えば，人物 x と人物 y に対して，“x AND y” をクエリーとしてヒット件数を求め，Jaccard 係数や Overlap 係数などの共起指標を計算し，関係のある／なしを決める．このとき，共起指標の値に対して，対象となる人物全体に一貫したしきい値を定め，例えば「共起ヒット件数が 50 件以上のペアにはエッジを張る」といった処理でネットワークを抽出する．本論文では，このような対象全体に一貫したしきい値を定めてエッジのあり／なしを決める方法を絶対的ルールによる関係抽出と呼ぶ．ノードの文脈に依存せず，共起指標の値としきい値によって絶対的な基準でエッジの有無が決定されるからである．

ところが，この絶対的ルールによる関係抽出には問題がある．様々な分野のアーティストが参加する国際的な展示会の場合，Web における名前の出現の頻度や共起の傾向にばらつきが大きく，アーティストによっては共起指標の値が大きく異なる．絶対的ルールによる関係抽出を適用した場合の典型的な問題は，ネットワークにおける中心的なノードのエッジはうまく抽出されるが，多くのノードは孤立してしまうことである．(逆に，孤立ノードを減らすようにしきい値を低く設定すると，ネットワーク全体でのエッジの数が多くなり過ぎる.）したがって，得られたネットワークが可視化や分析の目的に適さない場合もある．そもそも，社会学のネットワーク・クエスチョンでは，個々のノードに対して「あなたにとって重要な人は誰か」を問うことによってネットワークを抽出するものであった．そこに絶対的な基準はなく（人物 x にとっての人物 y の重要性と人物 z にとっての人物 w の重要性を比較しない），あくまでの個々のノードからみた相対的な重要性によってネットワークを抽出する．これを，本論文では相対的ルールによる関係抽出と呼ぶことにする．

図 1 に絶対的ルールによる関係抽出と相対的ルールによる関係抽出を模式的に示している．各エッジの太さは，共起指標により計量された関係の強さを表し，無向エッジである．仮にこの計量値をもとにエッジを 3 本だけ残すことを考える．絶対的ルールによる関係抽出では，A-B-C という中心的なアクター相互の関係がエッジとして選ばれる．共起指標の値が高いものから順に A-B, A-C, B-C であるからである．一方，相対的ルールによる関係抽出により，各ノードから見て最も共起指標が高いエッジを選ぶ処理を行うと A-B, A-C, D-E というエッジが採用される．このように絶対的ルールと相対的ルールでは，同じ計量値を用いても最終的に異なるネットワークが得られる．そして，この例では，相対的ルールにより得られたネットワークの方が，A の中心性の高さ，また 2 つのコミュニティ（A-B-C と D-E）の存在を適切に読み取れる．

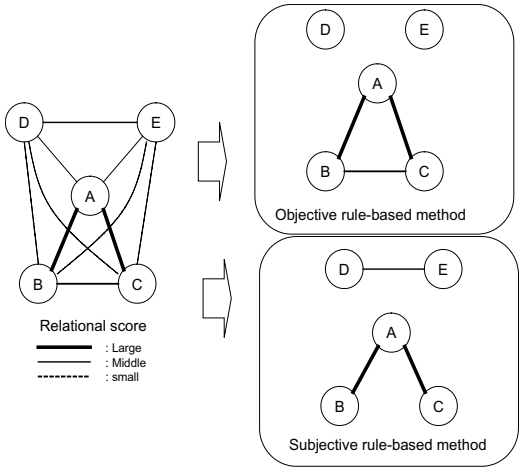


図 1　絶対的ルールによるネットワーク化と相対的ルールによるネットワーク化
Fig. 1　Extracted networks based on an objective rule and a subjective rule.

絶対的ルールと相対的ルールのどちらによってネットワークを抽出することが良いかは一概には言えない．例えば，ネットワークの中心的な人物を同定したい，主要なコミュニティを同定したいという場合には，絶対的ルールで問題ないであろう．しかし，周縁のコミュニティも適切に出したい，全体像を分かりやすく可視化したい，ネットワークをナビゲーションに用いたいといった場合には，相対的ルールの方が適している場合もある．したがって，目的に応じて，この二つの関係抽出を適切に組み合せることが重要である．

本論文では，絶対的ルールと相対的ルールを組み合せることで，目的により適したネットワークを抽出する方法を提案する．複数のパラメータを調整することで，この二つのルールを適切に融合する．評価実験により提案手法の有効性を示し，パラメータの性質について詳しく議論する．なお，本研究は，2005 年 9 月 28 日から 12 月 18 日にわたって横浜市で開催された横浜トリエンナーレ 2005(注2)のアーティストを対象に適用され，得られたネットワークは Web 上で閲覧・利用された．なお，本論文では,「社会的関係」という語を用いているが，対象とする関係が知り合いや好悪といった認知的な関係でなく，協働関係などの社会的な行為に基づく関係であるため，社会学における認知的な関係との対比の意味で用いている．

本論文は次のように構成される．**2.** では，Web から人間関係を抽出する基本的な考え方と従来手法，及びその問題点について説明する．**3.** では，提案手法のアルゴリズムについて詳しく述べるとともに，システム全体の流れについて説明する．**4.** では，評価実験と考察を通して提案手法の有効性を示す．**5.** で関連研究と本研究の位置づけなどについて議論し，**6.** でまとめを述べる．

2. Web から人間関係ネットワークの抽出

2.1 基本的な考え方

本論文で対象とする人間関係ネットワークにおいて，ノードは人であり，エッジは人同士の関係である．社会学ではそれぞれ，行為者と紐帯，もしくはアクター（actor）とタイ（tie）とよぶ．Web 上から人間関係を同定するということは，人間同士の実世界における関係の強さを，Web 空間における関係の強さから推定することである．そして，従来の研究では,「Web における名前の共起の強さは，その 2 人の実世界における関係の強さと高い相関がある」という仮説に基づいて関係の有無を判断している．

ここで，Web における名前の共起とは，同一の Web ページ上に名前が同時に出現することを指す．例えば，研究者の場合では，学会や研究会のプログラム，研究室のメンバーのページ，大学内の教官のメンバーリストなどのページに名前が多く共起するほど，２人のアクターの間には何らかの社会的関係が強い可能性が高いと推測できる．

2.2 従来手法：絶対的ルールによる関係抽出

Web における人物 x と y の共起の強さを計算する指標として，名前の共起頻度 ($|x \cap y|$) を直接用いる以外に，ダイス係数 ($2\frac{|x \cap y|}{|x|+|y|}$)，相互情報量 ($\log \frac{N|x \cap y|}{|x||y|}$)，コサイン類似度 ($\frac{|x \cap y|}{\sqrt{|x||y|}}$)，Jaccard 係数 ($\frac{|x \cap y|}{|x \cup y|}$)，Overlap 係数 ($\frac{|x \cap y|}{min(|x|,|y|)}$) などさまざまなものがある(注3) [7]. ただし，$|x|$，$|y|$，$|x \cap y|$，$|x \cup y|$ はそれぞれ，名前 x と名前 y の単独でのヒット件数，“x AND y” と “x OR y” でのヒット件数を表す．

Referral Web [5] や Flink [10] では，国際会議などに参加している有名な研究者を対象に，Jaccard 係数を用いて研究者同士の関係の強さを計算している．Polyphonet [8], [9] では，さまざまな指標を用いて人間関係の共起の強さを計算した場合について評価・考察を行い，Overlap 係数が人の協働関係の強さを表すのに最も適しているという知見を示している(注4)．原田らの NEXAS [3] では，与えられたトピックとそれに関するキーパーソンとの関連度を G スコアという指標を用いて計算している．いずれの場合も，ネットワーク全体で一貫したしきい値を設定し，共起指標がそれ以上であればエッジを張るという絶対的なルールに基づいてネットワークを構成する（図 2）．

2.3 従来手法の問題点

同じ分野の研究者（例えば，国内の人工知能の研究者や，国際的な Semantic Web の研究者）の場合，Web 上での名前の出現や共起の傾向に大きなばらつきはなく，絶対的ルールによる関係抽出によってうま

(注2)：www.yokohama2005.jp

(注3)：このような検索エンジンによる共起の指標が，どのような状況でどのような人間関係に対して有効であるかを明確にした研究は現在のところない．本手法で対象となる関係性は，多様な関係性の一部であることに注意頂きたい．

(注4)：Overlap 係数は，分母に関して min をとっており，ヒット件数の小さい方から見た距離感を表している．例えば，学生からみた先生との関係，有名でない人からみた有名な人との関係など，両方からみて強い方の関係をとる．

Input: a person name list L, and a threshold T
Output: a social network G

for each $x \in L$
  do set a node in G
done
for each $x \in L$ and $y \in L$
  do $r_{x,y} \leftarrow$ GoogleCooc(x,y)
done

/ Invent edges using subjective rule. */*
for each $x \in L$ and $y \in L$
  if $r_{x,y} > T$
    do set an edge between x and y in G
done
return(G)

* GoogleCooc returns the number of hits retrieved by a given query ("x AND y") using a search engine (Google).

図 2　従来手法：絶対的ルールによる関係抽出
Fig. 2　Algorithm of previous method.

くネットワークを抽出することができる．2 人の研究者が同じ研究会に参加する，同じプロジェクトに参加する，共著の論文を書くという行為によって，Web ページ上の氏名の共起となり，研究者の協働関係の強さがある程度適切に表される．これは研究者のコミュニティが比較的同質性が高い(注5)ことを暗黙的に利用している．

しかし，国際的に活動するアーティストの場合には状況は異なる．以下では横浜トリエンナーレに参加したアーティストを例に取る．例えば，日本の 2 人のアーティスト「安部泰輔」(x_1 とする）と「大榎淳」(y_1 とする）は横浜トリエンナーレ以外では関係がないが，Overlap 係数と Jaccard はそれぞれ $Overlap(x_1, y_1) = \frac{23}{min(113,397)} = 0.2035$, $Jaccard(x_1, y_1) = \frac{23}{960} = 0.024$ になる(注6)．ところが，スイスのアーティスト「Beat Streuli」(x_2 とする）とジャマイカのアーティスト「Nari Ward」(y_2 とする）の場合は，他に同じ展示会(注7) に参加したことがあるにも関わらず，$Overlap(x_2, y_2) = \frac{216}{min(89900,10400)} = 0.0208$, $Jaccard(x_2, y_2) = \frac{216}{175000} = 0.0009$ で共起関係が非常に弱く計算される．横浜トリエンナーレの参加アーティストにとって，他のすべてのアーティストと横浜トリエンナーレにおいて同じ展示会関係になりうるので，このような情報は価値が高くない．しかし，過去に同じ展示会に参加したことがあるという情報は，2 人のアーティストにの間にコミュニケーションがあったということであり，アーティスト間の繋がりを探る上で重要な情報である．しかし，共起関係の強さを絶対的に見た場合に，弱く計算された関係は取り逃すことが多い．これは一例であるが，一般に，様々な言語でコンテンツが書かれる Web では，同じ国（もしくは言語圏）の 2 人の名前が共起する可能性は，他の国（もしくは言語圏）の 2 人の名前が共起する可能性より高い．すなわち，国をまたがるアーティスト同士の関係は，同じ国のアーティスト同士より関係が弱く計算される場合が多い．同様に，現代美術と建築など異なる分野のアーティスト同士の関係は，同じ分野のアーティスト同士より弱い関係と計算される傾向にある．

また，最近結成されたアーティスト同士は，古くから結成されているアーティスト同士より，Web 上に名前が共起することが少ないので，共起指標の値が小さくなる．例えば，横浜トリエンナーレで「フライング・サーカス」という作品で初めて協力している「オン・ケンセン」と「ミール・ムハマド」という 2 名の共起頻度は 3 件しかなく，Overlap 係数と Jaccard 係数では 0.005 と 0.0454 に過ぎない．したがって，この 2 人の関係は他のグループ関係より非常に弱いと計量され，従来のしきい値を用いる方法ではこのような関係は抽出できない．

異なる国や異なる分野のアーティスト同士の関係，また新しく結成された関係などは，アーティスト同士のコミュニケーションを促進し，新しい出会いやコラボレーションを結びつけるという点では無視することのできない重要な社会的関係である．(社会学ではしばしば「弱い紐帯」と呼ばれる.）しかし，従来の絶対的ルールによる関係抽出では，このような関係を多く取り逃がし，結果的に多数の孤立ノードができる．したがって，このような「弱い社会的関係」をできるだけ精度よく抽出することができれば，コラボレーションやナビゲーションの支援に役立てることができると考えられる．

3. 提 案 手 法

3.1 相対的ルールによる関係抽出

提案手法では，ネットワーク全体では共起指標の値

(注5)：社会学では，同じ集合（set）に含まれるアクターのペアを同質（homogeneous）とし，異なる集合に含まれるアクターのペアを異質（heterogeneous）と呼んでいる [13]．本論文では，国や分野，趣味などが同じ（または，類似する）アクターの集合を，同質性が高いコミュニティとする．
(注6)：Google（www.google.co.jp）を用いる．検索結果は，2005 年 10 月 25 日時点のものであり，以下の例でも同様である．
(注7)：http://www.universes-in-universe.de/car/sharjah/2005/e-artist.htm

Input: a person name list L, and a threshold M
Output: a social network G

for each $x \in L$
  do set a node in G
done
for each $x \in L$ and $y \in L$
  do $r_{x,y} \leftarrow$ GoogleCooc(x,y)
done

/ Invent edges using objective rule. */*
for each $x \in L$
  do $Y_x \leftarrow$ ConnectedNodes(x), $\bar{Y}_x \leftarrow L \setminus Y_x$
  while $|Y_x| < M$ and $\bar{Y}_x \neq \phi$
    $y = \underset{y_j \in \bar{Y}_x}{\operatorname{argmax}} r_{x,y_j}$, $\bar{Y}_x \leftarrow \bar{Y}_x \setminus \{y\}$
    do set an edge between x and y in G,
      $Y_x \leftarrow Y_x \cup \{y\}$
  done
done
return(G)

* ConnectedNodes returns a node set connectted with x; $|X|$ returns the number of elements in a set X.

図 3　ネットワーク・クエスチョンの考え方：相対的ルールによる関係抽出
Fig. 3　Algorithm of Network Questionnaire.

が低くても，各ノードから見て値が高い場合にはエッジを張るという相対的ルールによる関係抽出を行う．図 3 にそのアルゴリズムを示す．それぞれの人にとって相対的に強い関係の人々を（M 人まで）抽出する．上述の例では,「Beat Streuli」と「Nari Ward」の関係は，共起が少なく絶対的ルールではエッジとして選ばれないが,「Beat Streuli」にとって「Nari Ward」が他の人より相対的に強い関係であるので，エッジでつながれることになる．

相対的なルールを用いる手法では，ネットワークを構成するすべてのアクターを同等に扱うが，これは適切でない場合もある．例えば，多くの人と関係を取り持つ人（コネクター）の場合，相対的ルールではうまく取り出せないこともある．また，研究者のコミュニティのネットワークを考えたときに，教授であっても学生であっても一様に同じだけのエッジの数が割り当てられるというのは，直観に合わない．つまり，アクターの関係的活動量(注8)そのものが大きく異なるときにも同様に扱ってしまうのが相対的ルールの欠点である．

したがって，本論文では相対的なルールだけを用いるのではなく，絶対的ルールと相対的ルールを組み合せたネットワーク抽出のアルゴリズムを構築する．

(注8)：次数はそのアクターの関係的活動量を示す．アクターが他のアクターと多くの紐帯を保持すればするほど次数が高く中心的である [16].

Input: a person name list L, and threshold set $< T, M >$
Output: a social network G

for each $x \in L$
  do set a node in G
done
for each $x \in L$ and $y \in L$
  do $r_{x,y} \leftarrow$ GoogleCooc(x,y)
done

/ First, invent edges using subjective rule.*/* ...(1)
for each $x \in L$ and $y \in L$
  if $r_{x,y} > T$
    do set an edge between x and y in G
done

/ Then, invent edges using objective rule.*/* ...(2)
for each $x \in L$
  do $Y_x \leftarrow$ ConnectedNodes(x), $\bar{Y}_x \leftarrow L \setminus Y_x$
  while $|Y_x| < M$ and $\bar{Y}_x \neq \phi$
    $y = \underset{y_j \in \bar{Y}_y}{\operatorname{argmax}} r_{x,y_j}$, $\bar{Y}_x \leftarrow \bar{Y}_x \setminus \{y\}$
    do set an edge between x and y in G,
      $Y_x \leftarrow Y_x \cup \{y\}$
  done
done
return(G)

図 4　提案手法：絶対的・相対的ルールによる関係抽出
Fig. 4　Algorithm of proposed method.

提案手法のアルゴリズムを図 4 に示す．これは，図 2 と図 3 を組み合せたものになっており，その処理は以下の通りである．まず，絶対的ルールによる関係抽出により，関係の強さが一定のしきい値 T 以上になるアクター同士をエッジでつなぐ．つぎに，相対的ルールによる関係抽出により，エッジの数が少ないアクターに対して，そのアクターのもつ相対的に強い関係のエッジを，エッジの数が M になるまで追加していく．したがって，提案手法は共起指標の絶対的な大きさと相対的な大きさの両方を考慮し，関係の活動量が多い人は多くのエッジをもつし，そうでない人でも少なくとも M 本のエッジは取り出されることになる．

3.2　構築したシステムのアルゴリズム

我々のシステムでは，提案手法を用いて Web からアーティストのネットワークを抽出する．その際，検索エンジンによる共起件数を用いた複数の共起指標を用いる．単独での共起指標でカバーできない事例を複数の指標を組み合わせることで，関係をよりロバストに抽出するためである．本システムでは共起頻度と

Overlap 係数を用いる(注9)．これまでの研究でも，抽出をよりロバストにするために複数の指標が用いられることがある．例えば，単独の名前のヒット件数が一定以上の研究者だけを対象としたり [8], [9]，平均のヒット件数以上の人だけを対象とする [10] などである．

Input: a person name list L, and threshold set $< T_{ov}$, T_{co}, M_1, $M_2 >$
Output: a social network G

for each $x \in L$
  do set a node in G
done
for each $x \in L$ and $y \in L$
  do $r^{ov}_{x,y} \leftarrow$ overlap(x,y), $r^{co}_{x,y} \leftarrow$ cooc(x,y)
done

/* *First, invent edges using subjective rule.**/ ...(1)
for each $x \in L$ and $y \in L$
  if ($r^{ov}_{x,y} > T_{ov}$ AND $r^{co}_{x,y} > T_{co}$) ..... (*RULE 1*)
    do set an edge between x and y in G
done

/* *Then, invent edges using objective rule.**/ ...(2)
for each $x \in L$
  do $Y_x \leftarrow$ ConnectedNodes(x),
    $\bar{Y}_x \leftarrow L \setminus Y_x$
  while $|Y_x| < M_1$ and $\bar{Y}_x \neq \phi$
    $y \leftarrow \underset{y_j \in \bar{Y}_x}{\text{argmax}}\ r^{ov}_{x,y}$, $\bar{Y}_x \leftarrow \bar{Y}_x \setminus \{y\}$
    if $r^{ov}_{x,y} > T_{ov}$ OR $r^{co}_{x,y} > T_{co}$ ..... (*RULE 2*)
      do set an edge between x and y in G,
        $Y_x \leftarrow Y_x \cup \{y\}$
  done
  $\bar{Y}'_x \leftarrow L \setminus Y_x$
  while $|Y_x| < M_2$ and $\bar{Y}'_x \neq \phi$
    $y \leftarrow \underset{y_k \in \bar{Y}'_x}{\text{argmax}}\ r^{ov}_{x,y}$, $\bar{Y}'_x \leftarrow \bar{Y}'_x \setminus \{y\}$
    if $r^{ov}_{x,y} > 0$ AND $r^{co}_{x,y} > 0$ ..... (*RULE 3*)
      do set an edge between x and y in G,
        $Y_x \leftarrow Y_x \cup \{y\}$
  done
done
return(G)

図 5　提案手法の応用：横浜トリエンナーレ 2005 に参加したアーティストの関係抽出．
Fig. 5　Detailed algorithm used in the Yokohama Triennale 2005.

(注9)：共起頻度は，単純に二つの集合の絶対的な重なりを表し，直感的に分かりやすい指標であるが，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．Overlap 係数は，ノードの数の少ない人からみた相対的な重なりを表し，人の協働関係を表すのに最も良い指標として知られている [9] が，単独でのヒット件数が非常に少ない人には特に高い値が出やすいという問題がある．これらの指標を組み合わせることで，一方でカバーできない事例をもう一方である程度補うことができる．それ以外にもさまざまな共起指標が提案されているが，どのような状況でどの指標が一番適切かに関しては，本論文では議論しない．

処理の詳細を図 5 に示す．まず，アーティストの名前のリスト L とパラメータ $< T_{ov}, T_{co}, M_1, M2 >$ を与える．各アーティストの氏名 $x \in L$，及び氏名のペア $(x, y) \in L$ に対して，検索エンジンのヒット件数を求める．x と y の共起関係の強さを，Overlap 係数と共起頻度の二つの指標で求め，$r^{ov}_{x,y}$ と $r^{co}_{x,y}$ とする．二つの指標がそれぞれのしきい値（T_{ov}, T_{co}）を満たせば，アーティスト同士をエッジでつなぐ．(これをルール 1 とする．) その結果，エッジの数が M_1 本より少ないアーティストに対しては，エッジを追加する（ルール 2）．$r^{ov}_{x,y}$ と $r^{co}_{x,y}$ のどちらかがしきい値を満たせば，追加するエッジの候補となる．さらに，それでもエッジの数が M_2 本より少ないアーティストに関しては，1 回でも共起のあるアーティストに対し，共起指標の高いものからエッジの数が M_2 になるまでつなぐ．

ここで，しきい値 T_{ov} と T_{co} は，前節で述べた絶対的ルールのパラメータであり，M_1 と M_2 は相対的ルールのパラメータである．抽出のルールをどのように組み合わせるかについては，何らかの学習手法で求めることも可能であるが，本論文では簡単のため上記の 3 つのルールを試行錯誤で作成し，用いた．なお，$M_1 = 0$, $M_2 = 0$ とした場合には，従来手法，すなわち絶対的なルールだけによるネットワーク抽出に相当する．

3.3　関係の判別とシステムの詳細

我々が構築したシステムでは，上述の方法によりネットワークを構成した後，各エッジに対して関係の種類（ラベル）を判別する．ここでは，アーティスト間の関係として 2 種類を考慮する．ひとつは，過去から現在まで，同じグループで作品の製作を行った「同グループ関係」である．もうひとつは，同じ展示会やイベントに出品したという「同展示会関係」である．これらの関係は，新しいユニットを結成する，作品を評価する，展示会を開催するなどの場面において，人間関係が少ながらず活用されることが多いだろう．いずれも検索した Web ページのテキストを分析することで取得する方法であり, [8] を参考にしている．すなわち，関係を表す属性を用いて学習データから関係を判別するルールを生成する．例えば，氏名とグループ関係を表す語群（例えば,「構成」「結成」「ユニット」など）が同じ文に出現するならば「同グループ関係」として判断する，氏名がリストやテーブルに現れて会議や展示会を表す語群（「展示会」「トリエンナーレ」「参加者」など）と同じページに出現するならば「同

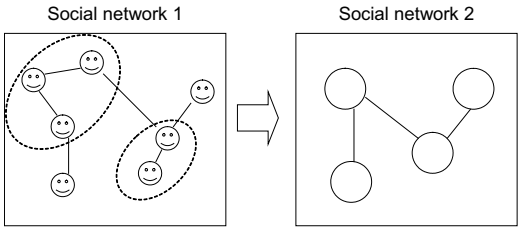


図 6　アクターの関係を用いたプロジェクト間関係の同定
Fig. 6　Identification of the relations among projects using actors' relations.

展示会関係」と判断する，など，簡単なものである．「同グループ関係」「同展示関係」のいずれとも判別されなかった関係は関係の根拠が見つからないものとして，エッジを削除する．

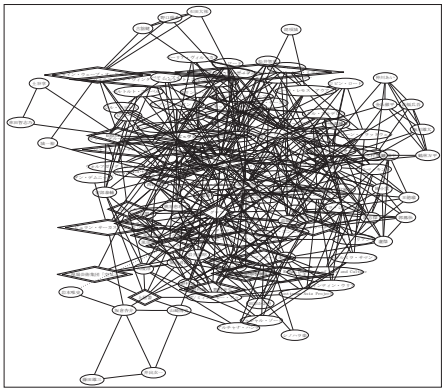
本研究では，関係の種類の判別は外部のモジュールとして捉え，従来手法，提案手法のいずれも同じものを用いている．次節の評価で示すように，全てのノードのペアに対してこのモジュールを適用しても精度が良くなるわけではない．ネットワーク抽出の段階を工夫し，弱い社会的関係であっても的確に捉えることで，同じモジュールを用いても抽出の精度を上げることができる．

なお，横浜トリエンナーレでは，ひとりのアーティストがノードになるとは限らず，アーティストが協力して作品を作る「プロジェクト」と呼ばれる単位がひとつのノードとなることもある．プロジェクトを含んだネットワークを構築するには，図 6 のように，プロジェクトを構成するアクターの関係を利用した．例えば，プロジェクト a のメンバーである「堀尾貞治」がプロジェクト b のメンバーである「米田知子」と「同展示会」関係があるとすると，プロジェクト a とプロジェクト b は「同展示会」関係であると判断する．

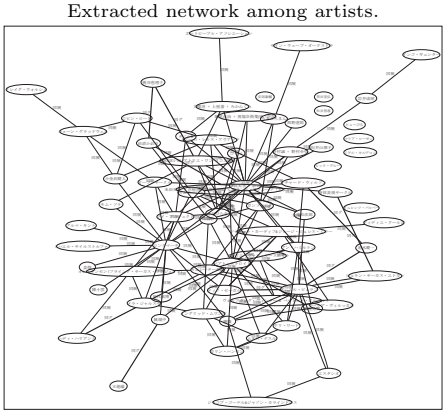
図 7 に横浜トリエンナーレ 2005 の参加アーティスト（132 人アーティスト，71 プロジェクト）に対して実際に得られたネットワークを示す．(a) はネットワークを抽出した段階であり，(b) は関係の種類を判別するモジュールを適用し，最終的に得られたネットワーク図である．

このネットワークを使ったアーティストの閲覧システムは，横浜トリエンナーレの開催期間中，トリエンナーレを支援する公式サポーターサイト（横浜シティアートネットワーク(注10)）から利用された(注11)．シス

(注10)：www.ycan.jp
(注11)：www.ycan.jp/archives/2005/11/polyphonet_arti.html 参照．



(a) 抽出されたアーティストのネットワーク
Extracted network among artists.



(b) 関係を同定し，絞り込まれたネットワーク
Improved network by identifying relations.

図 7　横浜トリエンナーレ 2005 のアーティストネットワーク
Fig. 7　Yokohama Triennale 2005 artist network.

テムのインタフェースを図 8 に示す．Flash で作られたこのインタフェースでは，調べたいアーティストの

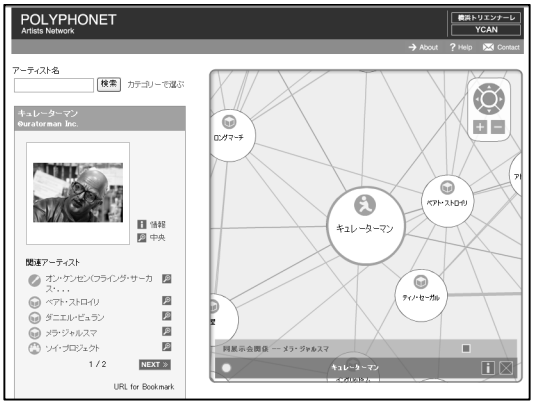


図 8　システムのインタフェース
Fig. 8　System interface.

名前を入力すると，そのアーティストが参加しているプロジェクトのノードが真ん中に移動し，そのプロジェクトと関係があるノードとのエッジがハイライトに表示される．エッジを順次たどっていくことで，関係の強いアーティストを次々と閲覧することができる．システムは http://mknet.polypho.net/tricosup/から利用可能である．

4. 提案手法の評価

提案手法の有効性を示すために，2種類のテストデータを作り，従来手法と提案手法による性能を比較した．まず，横浜トリエンナーレに参加したアーティストからランダムに 1000 ペアをサンプリングして，従来手法と提案手法による抽出結果を比較するとともに，各パラメータの性質を分析する．次に，人工知能学会全国大会 2006 に参加している 50 人の研究者を対象に，提案手法を適応することで，本手法の有効性を検証するとともに，パラメータ値の妥当性を示す．

4.1 アーティストネットワークの評価

テストデータは，132 人のアーティストに対する ${}_{132}C_2$ 組のペアから 1000 組をランダムに抽出し，人手で Web 上での関係性を調べた(注12)．今回の横浜トリエンナーレ以外の同展示会関係や同グループ関係などがあるかに着目し，正解を作った．テストデータには，146 組の関係のペアと 854 組の関係なしのペアが含まれている．そして，132 人のアーティストを入力として，4つのパラメータを変化させながら，それぞれのパラメータのセット $< T_{ov}, T_{co}, M_1, M_2 >$ で抽出されるネットワークの中で，上記の 1000 組のペアに対する適合率，再現率及び F 値による評価を行った．アクター同士の関係の強さは Overlap 係数を用いて計算して T_{ov} をしきい値とし，ヒット件数の制約は共起頻度を用いて，T_{co} で制約を与える(注13)．なお，従来手法で抽出されるネットワークは，パラメータが $< T_{ov}, T_{co}, 0, 0 >$ のネットワークに相当する．F 値は次の式で定義される．

$$F\text{ 値} = \frac{2 \times (\text{適合率}) \times (\text{再現率})}{(\text{適合率}) + (\text{再現率})}$$

表 1 は，従来手法を用いて，テストデータに対して最大の適合率，再現率，及び F 値が得られるときの各値とパラメータ T_{ov} 及び T_{co} の値を示している(注14)．最大の再現率を得るには T_{ov} や T_{co} を下限に設定すればよいので，$T_{ov} = 0$，$T_{co} = 0$ のとき再現率は 100%である．しかし，適合率は 14.6%と非常に低い．逆に適合率を最大にするには，共起の高いペアだけを選べばよいので，$T_{ov} = 0.24$，$T_{co} = 30$ と高く設定したときに，適合率は最大の 92.9%，再現率は 26.7%となる．両方のバランスを取る最も良いパラメータは，$T_{ov} = 0.05$，$T_{co} = 20$ であり，このとき F 値が最大の 0.50 となる．

一方，提案手法における結果を表 2 に示す．提案手法では4つのパラメータがあり，T_{ov} と T_{co} が表 1 と同じ値であるとした場合でも，M_1 と M_2 を適切に調整することで F 値が上がることを示している．全ての値を適切に調整すると，F 値は最大で 0.55 になる．括弧の中は，それぞれルール 1，ルール 2，ルール 3 から抽出されるエッジ数とその内の正解のエッジ数を示す．M_1 がゼロでない場合，ルール 2 により，ルール 1 でカバーできなかった関係を抽出することができる．例えば,「大榎淳」と「みかんぐみ」は同じ作品のために結成した「同グループ」関係があって，Overlap 係数と共起頻度はそれぞれ 0.162 と 162 件である．上記のパラメータ（$< 0.82, 20, 5, 1 >$）の場合，このような関係は，Overlap 係数ではカバーできないが共起頻度によりカバーすることができて，ルール 2 により抽出される．また M_2 がゼロでない場合，ルール 1 とルール 2 でもカバーできない関係をルール 3 により抽出することができる．例えば,「オン・ケンセン」と

(注12)：これらの 1000 組のペアをそれぞれクエリとして検索エンジンに入力し，ヒットされるページの中身を見て，関係性を判断する．同じグループやサーカス，パートナー，ユニットを結成した関係は,「同グループ関係」と判別し，同じ展示会やイベントに参加した場合は,「同展示会関係」と判別し，正解データとする．なお，横浜トリエンナーレ 2005 に出展したアーティストのペアが「同展示会関係」となるのは自明であるので，横浜トリエンナーレ 2005 における名前の共起は考慮しない．具体的には，次のような方法で関係の有無を判断した．まず，共起のヒット件数がゼロの場合は，そのペアは関係がないと判断する．また，共起のヒット件数がゼロでなくても，実際にページをみて関係がないと判断される場合には，関係がないとする．関係がないのに名前が共起する理由として，1 つは，同姓同名の問題がある．アーティストの名前には「graf」「SOI」「Open Circle」など，特殊なものが多いために起こる．もう 1 つの理由は，たまたま同じページの関係のない部分に名前が共起する場合で，例えば，アーティストや作品の紹介を中心とするサイトや個人のブログやニュース記事などで発生する．さらに関係が確実にないかどうかを調べるために，アーティストの名前単独で検索して，それぞれのページの中身を見て，同グループ関係や同展示会関係であるかを調べている．

(注13)：サーチエンジンは Google（www.google.co.jp）を用いる．なお，Google の検索結果は検索の度に変化する可能性があり，アルゴリズムの結果もその変化の影響を受けることがある．

(注14)：各パラメータは,T_{ov} は 0 から 1 まで 0.01 ずつ，T_{co} は 0 から 60 まで 5 ずつ，M_1 は 0 から 10 まで 1 ずつ，M_2 は 0 から M_1 までに変化させて検証した．以下でも同様である．

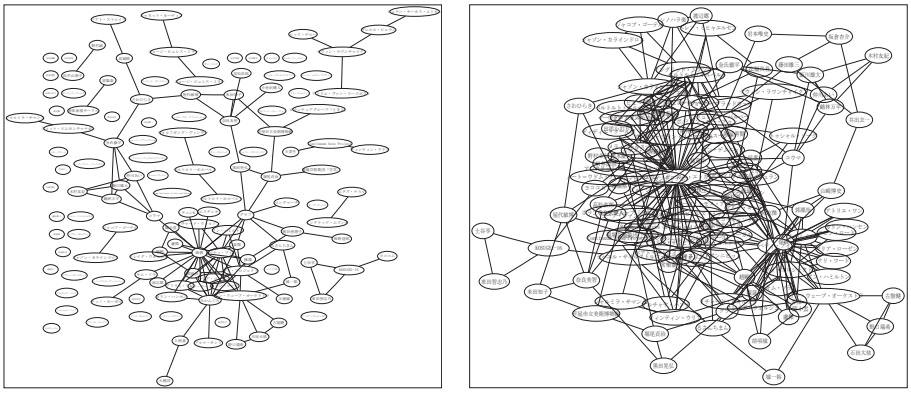
表 1 従来手法の適合率，再現率，F 値と各パラメータの値（アーティストのコミュニティ）
Table 1 Precision, recall, and F-value with parameters in the previous approach.

Cases	T_{ov}	T_{co}	P	R	F	#Extracted*	#Correct*
(a): Maximum Precision	0.24	30	92.9%	26.7%	0.41	42 (42,0,0)	39 (39,0,0)
(b): Maximum Recall	0	0	14.6%	100%	0.25	1000 (1000,0,0)	146 (146,0,0)
(c): Maximum F-value	0.05	20	76.4%	37.7%	0.50	72 (72,0,0)	55 (55,0,0)

*: Numbers in brackets are numbers of edges invented in $RULE1$, $RULE2$, and $RULE3$.

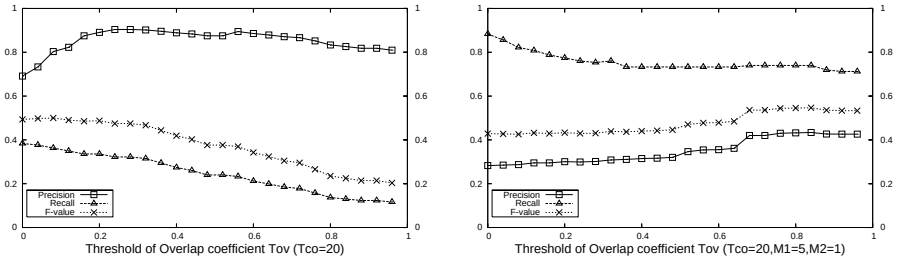
表 2 提案手法の適合率，再現率，F 値と各パラメータの値（アーティストのコミュニティ）
Table 2 Precision, recall, and F-value with parameters in the proposed approach.

Cases	T_{ov}	T_{co}	M_1	M_2	P	R	F	#Extracted	#Correct
Case (a')	0.24	30	3	2	34.4%	65.1%	0.45	277 (42,227,8)	95 (39,54,2)
Case (b')	0	0	0	0	14.6%	100%	0.25	1000 (1000,0,0)	146 (146,0,0)
Case (c')	0.05	20	1	0	55.4%	49.3%	0.52	130 (72,58,0)	72 (55,17,0)
(d'): Maximum F-value	0.82	20	5	1	43.4%	74.0%	0.55	249 (23,212,14)	108 (19,84,5)



(a) 従来手法 Previous approach. ($T_{ov} = 0.24$, $T_{co} = 30$)
(b) 提案手法 Proposed approach. ($T_{ov} = 0.24$, $T_{co} = 30$, $M_1 = 3$, $M_2 = 2$)

図 9 抽出されるネットワークの違い
Fig. 9 Difference of extracted networks.



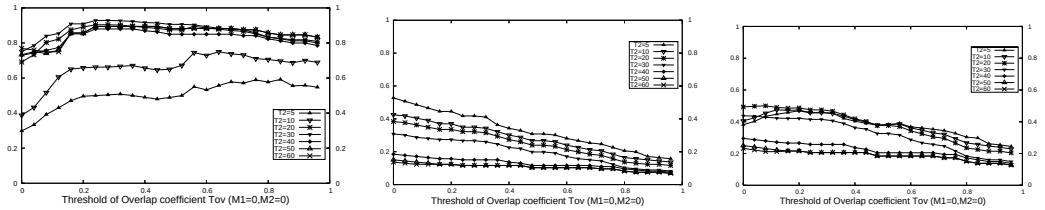
(a) 従来手法 Previous approach.
(b) 提案手法 Proposed approach.

図 10 T_{ov} を変化させたときの適合率，再現率，F 値の変化
Fig. 10 T_{ov} vs. precision, recall, and F-value.

「ミール・ムハマド」今回の展示会で初めて結成された関係なので Overlap 係数と共起頻度が 0.005 と 3 件しかないが，ほかに強い関係の人がいないので，ルール 3 によりカバーすることができる．

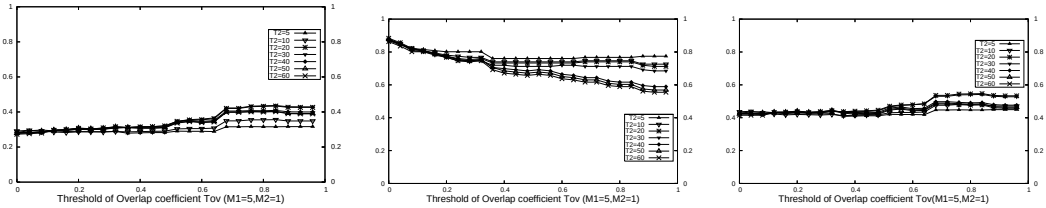
これらのパラメータを用いて構築されたネットワークを比較してみると，従来手法では図 9(a) に示すように多くの孤立ノードが生成されるのに対し，(b) の提案手法ではノードが孤立していない．

次に，この結果がどの程度ロバストかを，パラメータを変化させながら示す．図 10(a) と図 10(b) は，従来手法と提案手法に対し，T_{ov} を変化させたときに適合率，再現率，F 値がどのように変化するかを示し



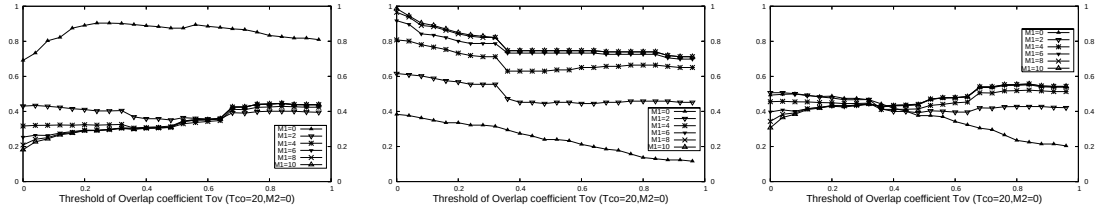
(a) 適合率 Precision　(b) 再現率 Recall　(c) F 値 F-value

図 11　T_{ov} と T_{co} を変化させたときの従来手法による性能の変化
Fig. 11　T_{ov} and T_{co} vs. performance in the previous approach.



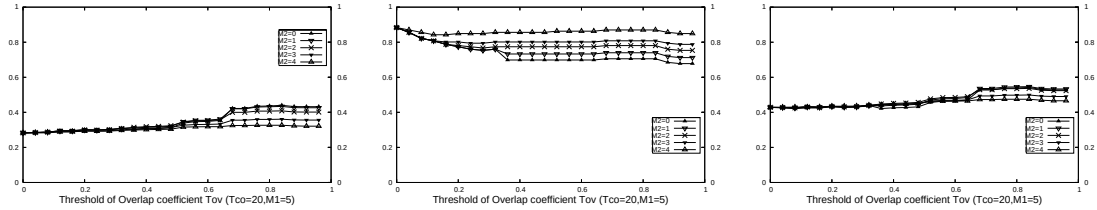
(a) 適合率 Precision　(b) 再現率 Recall　(c) F 値 F-value

図 12　T_{ov}，T_{co} を変化させたときの提案手法による性能の変化
Fig. 12　T_{ov} and T_{co} vs. performance in the proposed approach.



(a) 適合率 Precision　(b) 再現率 Recall　(c) F 値 F-value

図 13　M_1 を変化させたときの提案手法による性能の変化
Fig. 13　M_1 vs. performance in the proposed approach.



(a) 適合率 Precision　(b) 再現率 Recall　(c) F 値 F-value

図 14　M_2 を変化させたときの提案手法による性能の変化
Fig. 14　M_2 vs. performance in the proposed approach.

ている．(a) からは，2 節で従来手法の問題点として述べたように，T_{ov} を高く設定すると実際に存在する弱い社会的関係を取り逃がし再現率が低くなり，逆にしきい値を低く設定すると関係のないエッジが多く生成され適合率が下がる．一方の (b) では，$M_1 = 5$, $M_2 = 1$ のパラメータを加えることで，再現率が従来手法より高い値を保ち，T_{ov} によらず F 値が比較的安定していることが分かる．

また，図 11 と図 12 では，T_{ov} の変化，T_{co} の変化とともに (a) 適合率，(b) 再現率，(c)F 値がどのように変化するかを示している．従来手法の図 11 では，T_{ov}（横軸）や T_{co}（図中の線の種類）による値の変化が適合率，再現率，F 値ともに大きい（つまりグラフの変動が激しい）が，提案手法の図 12 では，いずれの場合でも，グラフがほぼ同じ値で安定していることが分かる．

図 13 と図 14 は，提案手法での M_1 と M_2 の変化による，(a) 適合率，(b) 再現率，(c)F 値の変化を示

している．M_1 と M_2 が大きくなるにしたがって，適合率は下がり，再現率は上がる．図 13(c) の F 値では，グラフの形が大きく変化する様子が読みとれ，特に M_1 と T_{ov} を適切に設定することが重要であることが分かる．つまり，Overlap 係数で全体に一貫したしきい値を設定しながら，かつ各ノードからの相対的に強い一定のエッジ数をうまく組み合わせると最も良い結果となる．

ここで示した結果は，提案手法が従来手法よりも良い F 値となり，提案手法が有効であることを示している．しかし，提案手法では従来手法よりアルゴリズムの自由なパラメータが増えているので，この結果はある程度当然の結果であるといえるかもしれない．しかし，ネットワーク全体に一貫したしきい値 (T_{ov}，T_{co}) を設定することと，各ノードから見たしきい値 (M_1，M_2) を設定することは，これまでにもネットワークを抽出する多くの研究で，無意識に混在されて用いられていた．本論文でこれをパラメータとして捉え，その性能評価となるデータを示せたことは，少なくとも Web や電子メール等から社会ネットワークを抽出するさまざまな研究に，重要な知見と示唆を与えるものである．

4.2 研究者ネットワークの評価

横浜トリエンナーレに参加したアーティスト以外のコミュニティに対しても，提案手法を適用することで，適切なパラメータを選び出すことができ，より正確にネットワークが得られることを示すため，人工知能学会第 20 回全国大会 (JSAI2006) に参加している研究者の中から，50 人（${}_{50}C_2 = 1225$ 組）の研究者を対象に，ネットワークの抽出を行う．提案の手法でネットワークを抽出するとともに，パラメータ値の決定方法についてその有効性を評価する．これらの 50 人の研究者間の協働関係（「共著関係」「同研究室関係」「同プロジェクト関係」「同発表関係」）の有無は，本人にアンケートの形で確認をとっている．

まず，50 人の研究者を入力として，4 つのパラメータを変化させながら，それぞれのパラメータのセットで抽出されるネットワークの中で，適合率，再現率及び F 値による評価を行った．表 3 は，従来手法を用いて，1225 組のペアに対して最大の評価となる場合の F 値とパラメータの値を表し，表 4 は，提案手法による結果を示す．従来手法で最大の F 値を出しているパラメータは，$< 0.2, 0 >$ の場合であり，研究者のコミュニティでは，Overlap 係数だけでも，比較的良い性能を出していることが分かる．また，同じ $< T_{ov}, T_{co} >$ を用いた場合，提案手法でも同様の F 値になる．ただ，提案手法で最大の F 値となる場合のパラメータは，$< 0.2, 20, 7, 0 >$ であり，従来手法より良いネットワークが得られることが分かる．

表 5　訓練データで F 値を最大にするパラメータをテストデータに適応した場合の性能
Table 5　Precision, recall, and F-value in the testing data with parameters that produced maximum F-values in the learning data.

(a) 従来手法 previous approach.

T_{ov}	T_{co}	F^L_{max}	P	R	F^T
0.18	0	0.66	84.6%	30.6%	0.45
0.8	0	0.66	100%	36.4%	0.53
0.12	0	0.60	60.0%	60.0%	0.60

(b) 提案手法 proposed approach.

T_{ov}	T_{co}	M_1	M_2	F^L_{max}	P	R	F^T
0.18	20	5	0	0.75	64.9%	66.7%	0.66
0.2	20	5	0	0.71	72.7%	72.7%	0.73
0.18	20	3	0	0.69	71.1%	67.5%	0.69

パラメータ値の決定方法の有効性を示すために，これらの 50 人の研究者のうち，40 人を訓練データとし残りの 10 人をテストデータとして，3 回の交差検定を行う．まず，訓練データで最大の F 値（F^L_{max} 値 ）を出力する場合のパラメータを，テストデータに適応した場合，抽出されるネットワークの適合率，再現率及び F 値（F^T値）を表 5 に示す．(a) は，絶対的指標だけをパラメータとした従来手法によるパラメータ値の検定で，(b) は提案手法の結果である．訓練データで最大の F 値を出しているときのパラメータは比較的安定している．また，いずれも提案手法の方がより安定して良い性能を示していることが分かる．

ネットワーク・クエスチョンの考えに基づいた相対的ルールのみ（図 3）でネットワークを生成した場合の性能を表 6 に示す．それぞれのノードに相対的に強い関係（ここでは，Overlap 係数の強い順に）のエッジを M 本まで追加した場合の適合率，再現率，及び F 値を比較すると，$M = 6$ の場合に最適なネットワーク（F 値=0.54）が得られる．しかし，いずれの場合も，提案手法で得られたネットワークより低い性能を示している．これは，相対的ルールに基づく方法ではいろいろな人と関係をもっている人の関係性が無視されてしまうという欠点を示している．

5. 関連研究と議論

どのような関係性をもつ人間関係を抽出するかに

表 3　従来手法の適合率，再現率，F値と各パラメータの値（研究者のコミュニティ）
Table 3　Precision, recall, and F-value with parameters in the previous approach.

Cases	T_{ov}	T_{co}	P	R	F	#Extracted*	#Correct*
(a): Maximum Precision	0.6	5	100%	5.98%	0.11	14 (14,0,0)	14 (14,0,0)
(b): Maximum Recall	0	0	19.1%	100%	0.32	1225 (1225,0,0)	234 (234,0,0)
(c): Maximum F-value	0.2	0	87.4%	47.4%	0.62	127 (127,0,0)	111 (111,0,0)

*: Numbers in brackets are numbers of edges invented in $RULE1$, $RULE2$, and $RULE3$.

表 4　提案手法の適合率，再現率，F値と各パラメータの値（研究者のコミュニティ）
Table 4　Precision, recall, and F-value with parameters in the proposed approach.

Cases	T_{ov}	T_{co}	M_1	M_2	P	R	F	#Extracted	#Corrct
Case (a')	0.6	5	3	3	31.7%	59.4%	0.41	438 (14,422,2)	139 (14,124,1)
Case (b')	0	0	0	0	19.1%	100%	0.32	1225 (1225,0,0)	234 (234,0,0)
Case (c')	0.2	0	0	0	87.4%	47.4%	0.62	127 (127,0,0)	111 (111,0,0)
(d'): Maximum F-value	0.2	20	7	0	68.0%	71.8%	0.70	247 (30,217,0)	168 (26,142,0)

表 6　相対的ルールによる関係抽出の結果
Table 6　Precision, recall, and F-value in the subjective rule.

M	P	R	F	#Extracted	#Correct
1	65.1%	12.0%	0.20	43	28
2	59.0%	19.7%	0.29	78	46
3	57.8%	28.6%	0.38	116	67
4	56.8%	37.6%	0.45	155	88
5	56.1%	47.4%	0.51	198	111
6	**53.8%**	**54.7%**	**0.54**	**238**	**128**
7	49.1%	58.1%	0.53	277	136
8	45.7%	61.5%	0.52	315	144
9	43.1%	64.1%	0.52	348	150
10	41.3%	67.5%	0.51	383	158

よって，ネットワークの最適なパラメータは異なる．パラメータを調整するには，ある程度の学習データを用意する必要がある．学習データをどのように用意するかに関しては，汎用的な手順がないが，例えば，横浜トリエンナーレのアーティスト同士の関係性を調べる際は，ホームページで公開されている同じプロジェクトに参加しているアーティストを正解データとして，初期段階のパラメータを定めた．また，システムのインタフェースにおいて，ユーザが抽出されたエッジの正誤を簡単に入力できる仕組みを採り入れることで，パラメータを自動的に学習・修正するシステムも可能であろう．

Web 上の情報から人間関係ネットワークを抽出する従来の手法では，これまで比較的同質なコミュニティを対象にしていた．接触頻度が多く，日常生活や研究活動において共通する部分が多い人間同士の関係を Web 上から見つけ，コアのメンバーについて正しく俯瞰を得たり，多くの人と連携するコネクターを見つけ出すことができる．社会学では，このような関係を「強い紐帯（Strong Ties）」と呼んでいる．しかし，強い紐帯で結ばれた人々はいろいろな側面において似ているので，情報のカバーしうる範囲が狭い．これに対し，接触頻度が低く，異なる分野や異なる社会圏の人同士を連結している「弱い紐帯（Weak Ties）」は，ブリッジのように情報伝達や社会統合で優れた力を発揮している [2]．様々な分野のアーティストが参加する国際的な展示会の場合，異なる国や分野のアーティスト同士の関係は弱い紐帯であると考えられ，その関係を見つけてつなぐことはコミュニケーションやコラボレーションの支援に重要である．本研究で提案しているネットワーク抽出手法は，従来方法と社会学におけるネットワーク・クエスチョンという典型的なデータ収集方法を統合するものであり，Web からの人間関係ネットワーク抽出をより多くの対象に適用可能にするために必要な技術である．

本研究で扱っている問題は一見単純なように思えるが，そもそもネットワークとは何かということに関わる重要な問題であると著者らは考えている．社会学におけるネットワーク分析は，ある現象をアクター自身の属性に帰着させるのではなく，その関係性に帰着させて説明しようというアプローチであった．しかし，関係性といっても，両者を含むイベント自身に意味がある場合，例えばメールのやりとりや電話でのやりとりといったこと自体を扱いたいという場合と，アクターから見た関係性の意義付けに意味がある場合，例えば誰が誰を信頼しているかや誰が一番好きかを扱いたい場合，この両者には大きな違いがある．ネットワークとは，そもそも現象の何をどのような目的で捉えようとするかに依存するものであり，ネットワークは存在するのではなく，むしろそう「見え」ているのである．本研究はこういったネットワークに関する本

質的な議論につながるものであると考えている．

6. むすび

本研究では，アーティストのような弱い社会的関係であっても，Web 上から適切にネットワークを抽出する手法について提案し，評価実験により提案手法の有効性と各パラメータの影響を示した．本手法で得られたアーティスト間のネットワークは，横浜トリエンナーレの開催期間中に Web サイト上で運用された．

今後は，様々な目的のネットワークにおいて，各パラメータはどのようなものが適切であるのか，またエッジのラベルを判別するモジュールの性質がどのようにネットワーク抽出の精度の向上につながるかといった研究をさらに進めていきたいと考えている．

文　献

[1] R. Bekkerman and A. McCallum, “Disambiguating web appearances of people in a social network,” Proc. 14th International World Wide Web Conference (WWW2005), pp.463–470, Chiba, Japan, 2005.

[2] M. Granovetter, “Strength of weak ties,” American Journal of Sociology, vol.78, pp.1360–1380, 1973.

[3] 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，vol.DBS-130/FI-71, 2003.

[4] 金光 淳，“社会ネットワーク分析の基礎 –社会的関係資本論にむけて–，” 勁草書店，2003.

[5] H. Kautz, B. Selman, and M. Shah, “Referral web: combining social networks and collaborative filtering,” Communications of the ACM, vol.40, no.3, pp.63–65, 1999.

[6] H. Kautz, B. Selman, and M. Shah, “The hidden web,” AI Magazine, vol18, no.2, pp.27–35, 1997.

[7] C.D. Manning and H. Schütze, “Foundations of statistical natural language processing,” The MIT Press, London, 2002.

[8] 松尾 豊, 友部 博教, 橋田 浩一, 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.1E, pp.46–56, 2005.

[9] Y. Matsuo, J. Mori, M. Hamasaki, K. Ishida, T. Nishimura, H. Takeda, K. Hasida, and M. Ishizuka, “POLYPHONET: an advanced social network extraction system,” Proc. 15th International World Wide Web Conference (WWW2006), pp.397–406, Edinburgh International Conference Centre, Scotland, 2006.

[10] P. Mika, “Flink: semantic web technology for the extraction and analysis of social networks,” Journal of Web Semantics, vol.3, no.2, pp.211–223, 2005.

[11] T. Miki, S. Nomura, and T. Ishida, “Semantic web link analysis to discover social relationship in academic communities,” IEEE/IPSJ Symposium on Applications and the Internet (SAINT-05), vol.00, no.0-7695-2262-9, pp.38–45, 2005.

[12] S. Staab, P. Domingos, P. Mika, J. Golbeck, L. Ding, T. Finin, A. Joshi, A. Nowak, and R. Vallacher, “Social networks applied,” IEEE Intelligent Systems, vol.20, no.1, pp.80–93, 2005.

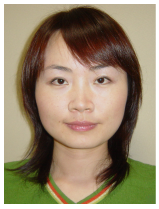
[13] S. Wasserman and K. Faust, “Social network analysis. methods and applications,” Cambridge University Press, Cambridge, 1994.

[14] B. Wellman, “The Global Village: Internet and Community,” University of Toronto, Idea&s - The Arts & Science Review, vol.1(1), pp.26–30, 2004.

[15] 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,

[16] 安田 雪，“実践ネットワーク分析，” 新曜社, 2001

（平成年月日受付，月日再受付）



金　英子

2001 年（中国上海）華東師範大学物理学部卒業．同年騰龍計算機軟件（上海）有限公司入社．2006 年東京大学大学院情報理工学系研究科修士課程終了．現在，同大学院博士課程在学中．Web マイニング，言語処理等に興味がある．人工知能学会，言語処理学会の各会員．

松尾　豊

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情報技術研究部門勤務，2005 年 10 月よりスタンフォード大学客員研究員．2007 年 10 東京大学工学系研究科総合研究機構准教授．人工知能，特に高次 Web マイニングに興味がある．人工知能学会，情報処理学会，AAAI の各会員．

石塚　満　（正員）

1971 年東京大学工学部電子卒，1976 年同大学院博士修了．工博．同年 NTT 入社，横須賀研究所勤務．1978 年東大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，2005 年同創造情報学専攻（電子情報学専攻兼任）．研究分野は人工知能，Web インテリジェンス，次世代 Web 情報基盤，生命的エージェントによるマルチモーダルメディア．IEEE, AAAI, 情報処理学会，人工知能学会（前会長），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．

Abstract Social network extraction from the Web is receiving much attention recently. This paper presents a new algorithm to extract a social network of artists. The algorithm can identify weak relationships among artists. We first describe the basic idea of extracting social networks from the Web, and then indicate that *objective rule-based methods* function ineffectively when applied to inhomogeneous communities. We propose a *subjective rule-based method* that is inspired by network questionnaires in social science. Furthermore, we propose to combine *objective and subjective rule-based methods*, which enables more exhaustive extraction than that under the previous method. We evaluate our method elaborately and demonstrated the effectiveness of our method. Our system was used at the International Triennale of Contemporary Art (Yokohama Triennale 2005) to facilitate navigation of artists' information.

Key words Web mining, social network, relation extraction, weak relationship