人間関係の重なりを持つコミュニティ構造の抽出

風間 一洋　佐藤 進也　斉藤 和巳　山田 武士

本論文では，人間関係のネットワークから活発な人間で構成されるコミュニティ構造を互いの重なりを許容しながら抽出する SR-2 法を提案する．本手法は，スペクトラルグラフ分析の一手法であり，ネットワーク構造中で他と重なりを持つような結合が密なコア部を抽出できる特徴を持つ SR 法を，特に共起ネットワークに対して，より詳細な分類ができるように変更したものである．この特性を調べるために，SR-2 法に加えて SR 法と k-クリークコミュニティ法を，実際の Web データから抽出した小規模な人間関係に適用して抽出されたノード集合を可視化すると共に，抽出性能を評価する．さらに，より大規模なネットワークとして論文の共著関係を取り上げ，各手法で抽出されるノード集合のサイズの分布を分析する．この結果，SR-2 法は，現実の人間の集まりに対応した妥当なコミュニティ構造を抽出できることを示す．

We present a SR-2 method to extract overlapping community structures from a human relationship network. It is a method of spectral graph analysis and is a variation of a SR method, which can extract a set of densely connected nodes from a network, changed for the more detailed extraction of co-occurrence networks. To investigate the characteristics of a SR method, a SR-2 method, and a k-community method, we visualized the results of extraction by their methods and evaluated the performance of extraction on a small human relationship network extracted from Web space. And we analyzed the distribution of node set sizes on a large co-authorship network. These results show that a SR-2 method can extract adequate community structures corresponding to real human communities.

1　はじめに

複雑ネットワーク研究の分野では，与えられたネットワークの構造を分析するために，ネットワークの結合の密な部分を抽出するさまざまな手法が提案されている．これらの手法の大部分が 1 つのノードは 1 つのノード集合にのみ属すると仮定しているが，現実には必ずしもそうとは限らない．たとえば，Web 空間においては単一の Web ページであっても複数のトピックを含むことも多く，人間関係においてはキーパーソンは多くの活動に関与する傾向があることから，人間関係のハブとなって多くの組織や人間を結びつけてしまう傾向が強い．現実のネットワークはこのような多重性を持つノードを含んでいるために，このような多重性を考慮しない既存の手法を用いると，うまく分離できずに非常に大きいノード集合として抽出されてしまう問題がある．

この問題を解決するために，ネットワークからノード集合を互いの重なりを許容しながら抽出する手法がいくつか提案されている．我々の SR(Spectral Relaxation) 法[12] は，スペクトラルグラフ分析の一手法であり，ネットワーク構造中で他と重なりを持つような結合が密なコア部を再帰的に抽出できる特徴を持つ．ただし，SR 法をブログのトラックバックネッ

Extraction of Overlapping Community Structures from Human Relationships.

Kazuhiro Kazama, Shin-ya Sato, 日本電信電話株式会社 NTT 未来ねっと研究所, NTT Network Innovation Laboratories, Nippon Telegram and Telephone Corporation.

Kazumi Saito, Takeshi Yamada, 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所, NTT Communication Science Laboratories, Nippon Telegram and Telephone Corporation.

コンピュータソフトウェア, Vol.24, No.1 (2007), pp.81–90.
[論文] 2006 年 2 月 20 日受付.

トワークに適用した時には良好な結果を得たが，人間関係ネットワークに適用した場合にはうまく分離できなかった．これは，対象とする人間関係ネットワークは共起関係に基づいて抽出されるのだが，SR 法がそのような特性を持つネットワークに適していなかったからだと思われる．

そこで，本論文では特に共起ネットワークの場合により詳細な分類ができるような変更を加えた SR-2 法を提案する．この特性を調べるために，SR-2 法に加えて同様に重なりを許容する抽出法である SR 法と k-クリークコミュニティ法を用いて，実際の Web データから抽出した小規模な人間関係に適用してノード集合を可視化すると共に，抽出性能を評価する．さらに，より大規模なネットワークとして論文の共著関係を取り上げ，各手法で抽出されるノード集合のサイズの分布を分析する．これらの結果から，SR-2 法は，現実の人間の集まりに対応した妥当なコミュニティ構造を抽出できることを示す．

2 関連研究

社会ネットワーク分析では，特に強い相互関係を持つ人々のグループに着目し，最大完全部分グラフである**クリーク**(clique) や，類似定義による部分グラフを分離するためのさまざまな研究が行なわれてきた [14]．しかし，ネットワークが巨大で密な場合や，自動抽出によって作成された場合などは，密な部分グラフが交差して多数存在するような複雑なネットワーク構造になり，適切な分離が困難になってしまう．

このような場合にも適切に分類できるようにするために，最近はノード集合の間の重なりを許容して抽出する方法が提案されている．たとえば，Palla らは，互いに隣接する k-クリーク (大きさ k の完全部分グラフ) の和集合を探索して **k-クリークコミュニティ**として抽出する手法を提案し [11]，このプログラムを CFinder として配布している†1．また，我々の提案する SR 法では，ネットワークの隣接行列の固有ベクトルに基づいてノードをランキングしてノード集合を求めた後に，既に抽出したコア部のエッジを削除する処理を再帰的に繰り返すことで，複数のノード集合を抽出する．SR 法は，単なる局所探索ではなくネットワーク全体から結合が強い部分を探索するために最大コアから順に求まり，同時にノードのランキングも決まること，さらに近似解法であるゆえに大規模ネットワークも高速に処理できるなどの利点がある．これに対して，Palla らの方法では k の値は経験的に決定しなければならないこと，また k の値が大きいクリークが存在すると計算時間が爆発的に増加するために，結合が密な大規模ネットワークには適さないという問題がある．

本手法は，ネットワークのクラスタリング (コミュニティ抽出) に関する研究とも関連がある．既存の多くのコミュニティ抽出法 [13] [3] [4] で，比較的密結合する 2 つの部分の間の隘路で分離する．これに対して，本手法は，結合が密なノード群のコア部を直接探索して，ノードをランキングする点と，多くの従来法のような排他的グラフ分割ではなく，ノード群の重複を許容したコア部の抽出を行なう点が異なる．

さらに，本手法は，スペクトラルグラフ理論 [2] とも密接な関係があり，本来の目的関数の探索を固有値問題として緩和して解いている点が共通する．ただし，代表的な normalized cut 法 [13] などがクラスタリング問題を扱うのに対して，本手法では抽出問題を扱う．一方，HITS [7] や PageRank [1] [10] のようなランキング手法もスペクトラルグラフ分析の一種であるが，本手法では，式 1 のように平均エッジ数の概念を導入することでコア部抽出を可能にするとともに，コア部に含まれるエッジを削除して再帰的に別のコア部抽出を実現している点が異なる．また，固有ベクトルを複数求める点では，主成分分析 PCA(Principal Component Analysis) [5] や潜在意味解析 LSA(Latent Semantic Analysis) [9] に類似し，本手法は平均エッジ数の多いコア部を抽出する射影法と見ることができるが，求める固有ベクトル群に直交制約はないことから，自由度が高く，より精緻な射影軸が求まることが期待できる．

最後に，与えられたネットワークのコア抽出問題に関しては，Web Trawling [8] などの先行研究がある．Web Trawling では，離散探索問題として多様な

†1 http://www.cfinder.org/

プルーニング手法を駆使しているのに対して，本手法は，連続量で最適解が求まる緩和問題を利用している点，平均エッジ数が多い順にコアを抽出できる点が異なる．

3 重なりを持つノード集合の抽出

本論文では，SR 法[12] を，人間関係のような共起関係から導出されるネットワーク用に拡張した SR-2 法を使用する．SR 法では，Web のハイパーリンクのように比較的粗であるネットワーク構造で，エッジが密集している部分には類似ノードが集まっていると仮定して，そのコア部を抽出する．この方法の特徴は，隣接行列の固有ベクトルを求めてノードのランキングを行なうことで結合が密なコア部を抽出すること，および既に抽出したコア部の内部エッジだけを削除することで別のコア部も再帰的に抽出することである．これにより，ネットワーク中の重なりを持つ密なノード集合の抽出が可能である．

以下で，本手法のアルゴリズムについて述べる．

3.1 基本設定

与えられた連結であるネットワークの全ノード集合を $S = \{1, \cdots, N\}$ とし，その隣接行列を $\mathbf{A}$ とする．隣接行列の第 (i, j) 成分の $A(i, j)$ は，ノード i と j 間にエッジがあれば 1 に，なければ 0 に設定する．なお簡略化のため，自己エッジなし $(A(i, i) = 0)$ の無向グラフ $(A(i, j) = A(j, i))$ として扱う．

ノード集合 $C \subset S$ に対して，平均エッジ数は以下で定義できる．

$$G(C) = \frac{1}{2} \sum_{i \in C} \sum_{j \in C} \frac{A(i, j)}{|C|}. \qquad (1)$$

$|C|$ は C に含まれるノード数を表す．

既に述べたように，エッジが密集している部分には類似ノードが集まっていると仮定して，式 1 を最大にするノード集合 C の探索問題を解く．ただし，一般にノード数の多い大規模ネットワークでは組合せ爆発が起こりやすいので，緩和問題を最適に解くアプローチを取る．

3.2 抽出アルゴリズム

本抽出アルゴリズムを以下に示す．

E1. ネットワークを連結成分に分解し，ノード数が λ より大きい連結成分を抽出する;

E2. 連結成分を 1 つ取り出し，存在する場合は S とし，存在しなければ終了する;

E3. S の最大固有ベクトル $\mathbf{q}^*_m$ を求める;

E4. $\mathbf{q}^*_m$ を量子化して，ノード集合 $C_m(k^*)$ を求める;

E5. $i, j \in C_m(k^*)(= C^*_m)$ ならば $A(i, j) = 0$ に設定する．

E6. **E1** を再帰的に呼び出して，変更された $\mathbf{A}$ を処理する;

E7. **E2** に戻る;

ここで，最小ノード数 λ は，たとえばノード数が 2 のような小さな連結成分を除去するために用いられる．

最終的に結果は $C^*_1, \cdots C^*_M$ として求まる．抽出されるノード集合の数 M は自動的に決定される．なお，**E5** においては，抽出されたノード集合 C^*_m の内部のエッジだけを無効化し，外部へのエッジはそのままにしていることに注意すること．

以降で，さらに **E3** と **E4** の詳しい手順を述べる．

3.3 最大固有ベクトルの計算 (E3)

ノード集合 C に対して，N 次元ベクトル $\mathbf{q}$ を，$i \in C$ ならば $q_i = 1$，さもなければ $q_i = 0$ と定義すると，式 1 は以下のように書き換えられる．

$$G(\mathbf{q}) = \frac{1}{2} \frac{\mathbf{q}^T \mathbf{A} \mathbf{q}}{\mathbf{q}^T \mathbf{q}}. \qquad (2)$$

ただし，$\mathbf{q}^T$ はベクトル $\mathbf{q}$ の転置を表す．ここで，ベクトル $\mathbf{q}$ の各要素に対して連続値まで許容すれば，式 2 の右辺は Rayleigh 商に他ならない．よって，$G(\mathbf{q})$ の最大値は，行列 $\mathbf{A}$ で固有値を最大にする固有ベクトル $\mathbf{q}^*$ で与えられる．

$\mathbf{q}^*$ は，パワー法に基づいて以下のように求める．

F1. $t = 1$, $\mathbf{q}^{(0)} = (1, \cdots, 1)^T$ と初期化する;

F2. $\widetilde{\mathbf{q}} = \mathbf{A}\mathbf{q}^{(t-1)}$, $\mathbf{q}^{(t)} = \widetilde{\mathbf{q}} / \max_i \widetilde{q}(i)$ を求める;

F3. $\max_i |q^{(t)}(i) - q^{(t-1)}(i)| < \epsilon$ なら反復を終了する;

F4. $t = t + 1$ として **F2** に戻る;

ここで，ϵ は終了条件を制御する正の実数であり，反復終了後に $\mathbf{q}^* = \mathbf{q}^{(t)}$ として結果が求まる．

明らかに，$\mathbf{A}$ と $\mathbf{q}^{(0)}$ の全要素が非負のため，任意の反復で $\widetilde{\mathbf{q}}$ の各要素は非負となる．さらに，**F2** でスケーリングを施すことで，$0 \leq q^{(t)}(i) \leq 1$ が保証される．つまり、上記アルゴリズムで固有値最大の固有ベクトルが求まり，基本設定の妥当な緩和問題を最適に解いていると言える．

3.4 固有ベクトルの量子化 (E4)

本来は $\mathbf{q}$ は C に属するかどうかを示すバイナリベクトルであるが，本手法では緩和問題として解いているために，得られた固有ベクトル $\mathbf{q}^*$ は連続値となり，これは C への帰属度に対応すると解釈できる．そこで， $\mathbf{q}^*$ の要素の大小に基づき各ノードをランキングし，C への帰属度が高いものから順に k^* 個ノードを選択し，C の近似として $C(k^*)$ を構成することを考える．これは，$\mathbf{q}^*$ の要素をある閾値でバイナリに量子化することに対応する．

まず，$\mathbf{q}^*$ の要素の大小に基づき各ノードをランキングすれば，リスト $R = [r(1), \cdots, r(N)]$ が定まる．ここで，$r(i)$ はランク i に対して元のノード番号を与える関数で，$q^*(r(i)) \geq q^*(r(i+1))$ の関係を満たす．なお，タイブレークは任意に行なう．リスト $R = [r(1), \cdots, r(N)]$ の上位 k 個のノード集合

$$C(k) = \{r(i) : i \leq k\}. \tag{3}$$

を考えれば，その平均エッジ数は以下で求められる．

$$G(k) = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \frac{A(r(i), r(j))}{k}. \tag{4}$$

本手法では，式 1 を直接解く代わりに，k を増やした時の式 4 の最初のピークになる k^* を探索して，ノード集合 $C(k^*)$ を求める．k^* を効率良く探索するために，式 4 の定義より導ける以下の漸化式を利用する．

$$G(k+1) = G(k) + \frac{\Delta(k+1) - G(k)}{k+1}. \tag{5}$$

$\Delta(k+1)$ はノード $r(k+1)$ を加えたことによるエッジ数の増分であり，以下のように計算できる．

$$\Delta(k+1) = \sum_{j=1}^{k} A(r(j), r(k+1)). \tag{6}$$

一方，定義より $G(1) = 0$ である．

そこで，上記手順をまとめれば以下となる．

G1. $\mathbf{q}^*$ の要素をソートし，ランク関数 $r(i)$ を求める;

G2. $k = 1, G(1) = 0$;

G3. $G(k+1)$ を式 5 と式 6 で求める;

G4. $G(k+1) < G(k)$ なら $k^* = k$ として $C(k^*)$ を出力して終了する;

G5. $k = k + 1$ し，**G2** に戻る;

3.5 改良点

SR-2 法では，オリジナルの SR 法[12] に，次の 2 つの改良を施した．

SR 法では，ネットワークが連結でないと、必ずしもパワー法が最大固有値に収束する保証がない．しかし，本論文で扱うネットワークは必ずしも連結であるとは限らないので，一旦連結成分ごとに分割してから抽出している．

次に，固有ベクトルを量子化する時に，式 4 の最大値ではなく，最初のピークを探索するようにした．これは，比較的本質的な変更である．本論文で対象にした共起関係から導出されるネットワークでは，[12] で扱ったトラックバックネットワークとは異なり，同時に出現する複数のノード同士が互いに結びつくためにノード集合同士の境界がより曖昧であり，SR 法では期待するよりも大きなノード集合を抽出する傾向があった．そこで，エッジ数は比較的多くても括れている部分で切断して別のノード集合として分離して，より詳細な抽出ができるように変更した．

4 実験

4.1 実験用データ

本論文では，実験用データとして，2 種類の異なる手法で抽出された，大きさが異なる人間関係ネットワークのデータを使用した．

1 つは，別の論文[6] のシステムで抽出した，Web 上の特定のトピックに関する影響力がある人物達の人間関係ネットワークであり，そのトピックをよく表す検索語で検索して得られた検索結果の上位を選択し，それらの Web ページ上の出現位置を考慮した人名の

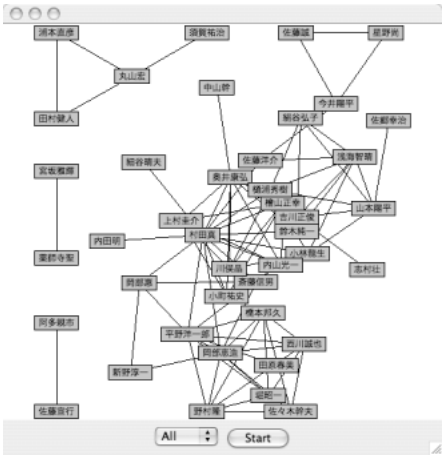

図 1 検索語 “XML” で得られた人間関係

共起関係を解析し，さらに人名の出現サーバ数を用いて影響力のある人名だけに絞り込んだ．実際に用いた検索語は “XML”，各パラメータは対象検索結果数 $n = 1000$，人名を共起と判定するページ内の距離の閾値 $d = 3$, 人名抽出に使用する閾値の人名の出現サーバ数 $e = 2$ である．特にこの検索語を選んだ理由は，活動が活発な分野であると共に，一部の著名人がさまざまな団体で活動しているために，複雑なネットワークが得られるからである．検索対象データは，jp ドメイン内，および jp ドメインから日本語のアンカーテキストでリンクされている HTML ファイルであり，収集時期は 2003 年 7 月，収集ページ数は 52,302,804 ページである．ノード数は 42 個，エッジ数は 83 本，平均エッジ数は 1.98 本である．この人名の抽出精度は 90.5%(人名の抽出ミスが 3 件，関連する分野だが直接の貢献がない場合が 1 件)，人間関係の抽出精度は 90.4%(人名の抽出ミス関連が 6 件，直接関係のない場合が 1 件，同一ページの別の書籍の著者の場合が 1 件) である．実験には，修正を加えずにそのまま使用した．図 1 に，ばねモデルで可視化した結果を示す．この図から，4 つの連結成分があることがわかる．

もう一つは，CFinder にも添付されている，Los Alamos cond-mat のアーカイブから抽出した共著者ネットワークである．これは 1998 年 4 月から 2004 年 2 月までの記事から抽出し，さらに n 人の著者の組の間のエッジの重みの値を $1/(n-1)$ として閾値 1 以下のエッジを除去して作成されている．ノード数は 16,662 個，エッジ数は 22,446 本，平均エッジ数は 1.35 本である．

4.2 抽出結果の可視化

まず，手法による抽出結果の違いを調べるために，小規模な Web 上の人間関係ネットワークに SR 法，SR-2 法，k-クリークコミュニティ法を適用し，その抽出結果の全ノード集合を可視化した．ただし，SR 法，SR-2 法では，可能な限りすべてのノード集合を抽出するように，抽出ノード集合数 N を指定せずに自動停止させ，かつ $\lambda = 1$ とした．また，k-クリークコミュニティ法では一番良い性能が得られた $k = 4$ を用いた．なお，k の値に伴う抽出結果の変化については，後述する．SR 法，SR-2 法，k-クリークコミュニティ法による可視化結果を，それぞれ図 2，図 3，図 4 に示す．なお，各図は図 1 のネットワーク構造の中で，各ノード集合に含まれるノードとエッジだけを強調表示している．

SR 法，SR-2 法，k-クリークコミュニティ法によって，それぞれ 16 個，11 個，4 個のノード集合が抽出された．便宜上，図 2，図 3，図 4 で示す各ノード集合群に対して，左から右，さらに上から下の順に，ノード集合 1，2，…と呼ぶことにする．

抽出されたノード集合の大きさに着目すると，SR 法の場合には最初にノード数が 15 個と大きいノード集合が抽出される傾向があるが，SR-2 法と k-クリークコミュニティ法ではそれが中規模の複数のノード集合に分割して抽出されている．抽出されたノード集合の形状に着目すると，SR-2 法と k-クリークコミュニティ法は完全グラフに近い形状を持つのに対して，SR 法は必ずしもそうとは限らない．なお，SR 法と SR-2 法では，各連結成分に対して，後の方で得られるノード集合はノード数も平均リンク数も小さい．この理由は，SR 法では，エッジ密度の少ないところではノード集合が抽出できたとしても固有値の差が小さいからである．このために，平均エッジ数が極端に低いノード集合は，結果として必ずしも信頼できるとは限らない．

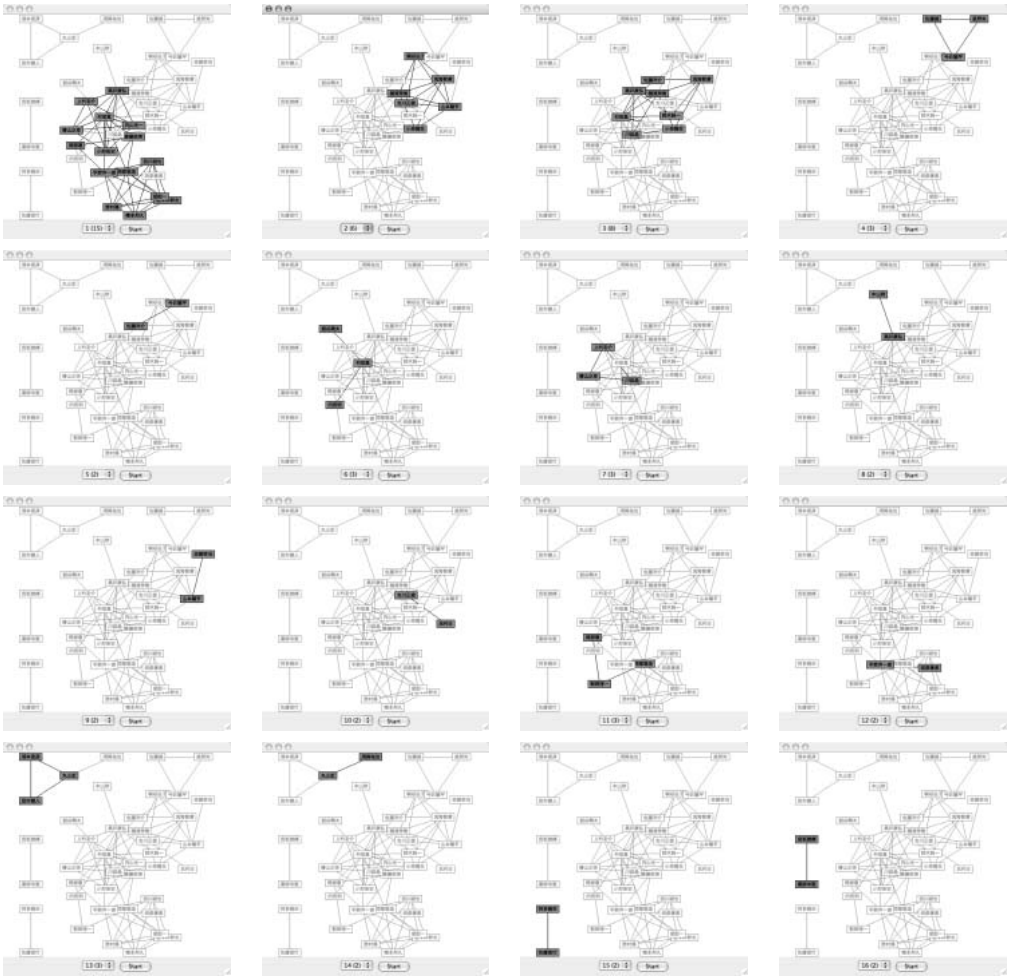

図 2　抽出されたノード集合 (SR 法)

4.3　抽出結果の妥当性の分析

次に，各手法の抽出結果にどのような意味的な違いがあるかを知るために，現実の人間の集まりと抽出されたノード集合との対応関係について調べ，妥当性を評価した．ただし，SR 法と SR-2 法では，最大連結成分の中で平均エッジ数が 1 以下のノード集合は意味がないものとして除外した．この結果，SR 法，SR-2 法，k-クリークコミュニティ法で評価対象となるノード集合数は，図 2 では 1〜3 の 3 個，図 3 では 1〜5 の 5 個，図 4 では 1〜4 の 4 個となった．

まず，XML で検索した時の上位 1,000 件の検索結果から，国内の最大の XML 開発者会議である XML 開発者の日，JIS などの XML 関連の標準化，XML 関係の業界団体である XML Publishing Forum，情報処理学会，XML の資格制度である XML マスターという 5 つの課題に対して，発表者，主催者，著者などの具体的な役割で関係している人物群を抽出し，それぞれ A, B, C, D, E という 5 つの評価用の人名集合を作成した．ここで，本手法で得られる人間関係は，単一の Web ページから求めているのではなく，さまざまな Web ページに登場する複数の断片的な人間関係から合成されていることに注意されたい．たとえば,「XML 開発者の日」では，開催時期の異なる複数のプログラムが,「標準化」では「拡張可能なマーク付け言語 (XML) 1.0」(XML の仕様) や「XML 日本語プロファイル」などの複数の仕様が対象となって

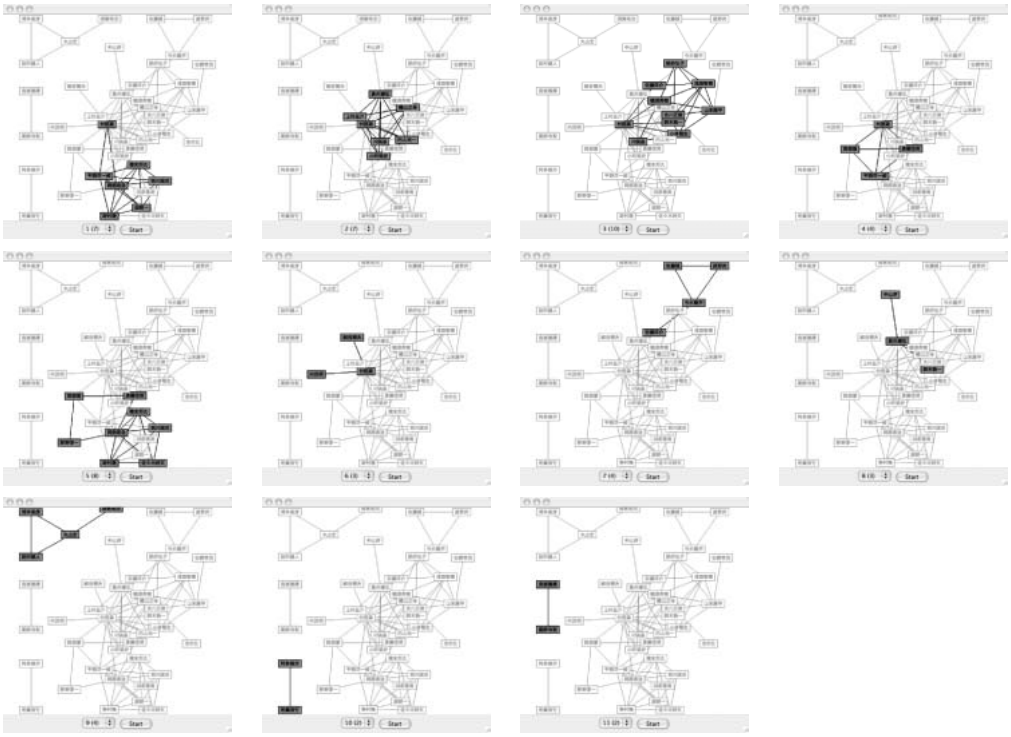

図 3 抽出されたノード集合 (SR-2 法)

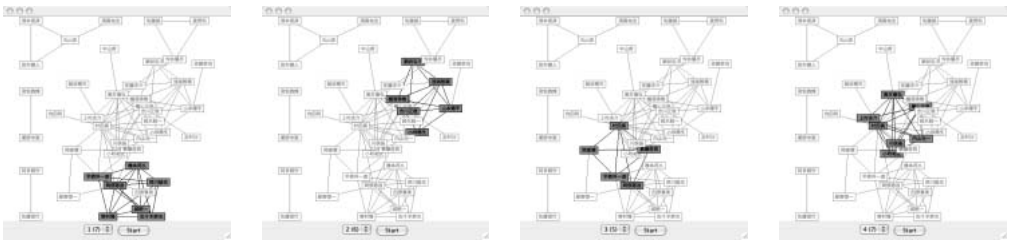

図 4 抽出されたノード集合 (k-クリークコミュニティ法)

いる．

なお，この 5 つの人名集合は，すべて図 1 ではほぼ中央に位置する一番大きなノード集合に含まれている．このように異なる人間の集まりが複雑に結合している理由は，日本における XML 関係のアクティブな活動は専門家の協調によって支えられているからであると推測している．ただし，XML は現在の Web システムや企業システムを支える重要な基本技術となって業界に広く普及しており，さらに同じ XML の専門家でも互いの詳細な専門や志向が異なるために，このような形状になると考えられる．分野によっては，必ずしもこのような形状になるとは限らないので，興味を持たれた場合は別の論文[6] も参照して頂きたい．また，他の図 1 のノード集合は，2 つが書籍に関連し，残りの 1 つが企業の経営陣の XML 戦略に関係していた．このように，専門家であっても，互いに協調して活動するタイプと，比較的独立に活動するタイプに分類でき，それに応じて得られるノード集合の形状や性質が異なってくる．

この間の対応関係を調べるために，各人名集合に

表 1　人間の集まりとノード集合の関係

SR 法					
No.	1	2	3		
ノード数	15	6	8		
A(11)	2	**6**	6		
B(9)	**7**	0	0		
C(7)	**6**	0	3		
D(7)	1	3	**4**		
E(5)	**5**	0	1		
SR-2 法					
No.	1	2	3	4	5
ノード数	7	7	10	4	8
A(11)	1	3	**9**	1	0
B(9)	1	**7**	2	1	0
C(7)	**6**	1	0	1	5
D(7)	1	1	**6**	1	0
E(5)	3	1	1	**4**	3
k-クリークコミュニティ法					
No.	1	2	3	4	
ノード数	7	6	5	7	
A(11)	0	**6**	1	3	
B(9)	**7**	0	2	0	
C(7)	0	0	1	**7**	
D(7)	0	**3**	1	1	
E(5)	2	0	**5**	1	

ノード集合の人名がどの程度含まれるかを調べて表1に示す．この表の一番上の行の数値はノード集合の番号であり，その下の行はノード集合に含まれるノード数である．さらに，各課題に対して，ノード集合ごとに合致する人名の数を示す．なお，太字で示した箇所は，その課題に対して，そのノード集合が一番よく適合していることを表す．

この結果から，抽出されたノード集合は，各課題の人名集合に比較的よく対応していることがわかる．ただし，抽出された 1 つのノード集合に，複数の人名集合が対応していることがある．この原因は，1 つは SR 法の 1 のように，抽出されたノード集合が大きすぎる場合，もう 1 つは人名集合 A と D のように，お互いが重なりが大きい場合がある．

次に，抽出性能について調べる．精度 (precision)P と再現率 (recall)R は，次のように定義される．

$$P = \frac{w}{w+x} \tag{7}$$

$$R = \frac{w}{w+y} \tag{8}$$

ここで，w は抽出された適合ノード数，x は抽出された非適合ノード数，y は抽出されなかった適合ノード数を示す．この結果を表 2 に示す．

表 2　精度と再現率

		SR 法	SR-2 法	k-クリーク コミュニティ法
P	A	1	0.9	1
	B	0.467	1	1
	C	0.4	0.857	1
	D	0.5	0.6	0.5
	E	0.333	1	1
R	A	0.545	0.818	0.545
	B	0.778	0.778	0.778
	C	0.857	0.857	1
	D	0.571	0.857	0.429
	E	1	0.8	1

この結果から，精度に関しては SR 法は低いことがわかる．これは SR 法は比較的大きなノード集合を抽出する特性のために，このような目的には適していないからであると考えられる．この点に関しては，SR-2 法は比較的良好な結果を得ている．精度においては，k-クリークコミュニティ法が一番優れている．ただし，再現率においては，特に大きな人名集合を対象とした場合の性能が悪い．k-クリークコミュニティ法では，どのようなノード集合が取り出されるかは k に依存するが，k はネットワーク全体に一律に適用されてしまうために，複雑な形状のノード集合を取り出すのは難しいと考えられる．

なお，どの手法でも精度が再現率に対して相対的に高い傾向がある．これは，どれもコア部だけを対象とするために，他との関係が弱いノードは除外されてしまうからである．

4.4　抽出ノード集合サイズの分析

次に，SR 法，SR-2 法，k-クリークコミュニティ法の巨視的な抽出の性質の違いを調べるために，より大きい Los Alamos cond-mat のアーカイブから抽出した共著者ネットワークに適用した．なお，SR 法，SR-2 法では $\lambda = 3$，k-クリークコミュニティ法では $k = 4$ として，どちらもサイズが 4 以上のノード集合を抽出対象とした．この設定で抽出されたノード集合の総数は，それぞれ 995 個，1801 個，319 個であり，ノード集合の最大サイズは 2052，58，13 である．各

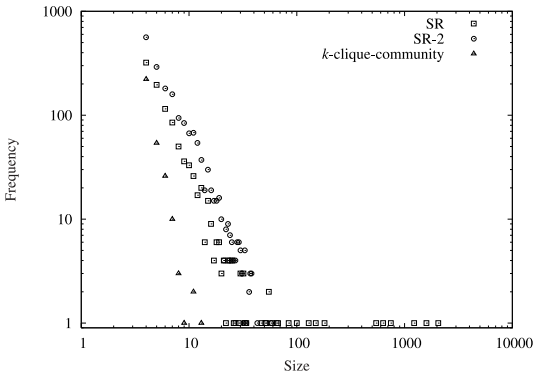

図 5 ノード集合のサイズ分布

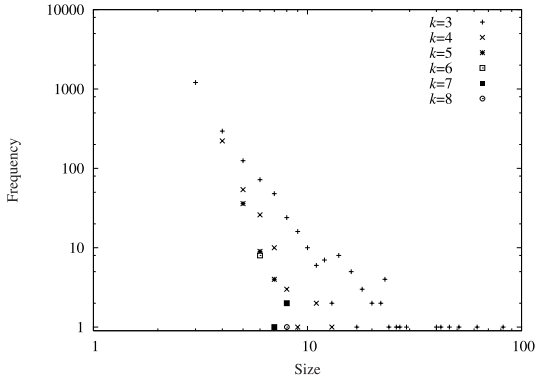

図 6 k によるサイズ分布の違い

手法ごとのノード集合のサイズの分布を，図5に示す．

この結果から，得られるノード集合のサイズは，k-クリークコミュニティ法が一番小さく，SR-2 法，SR 法の順で大きくなっていることがわかるが，これは前述の結果とも一致する．k-クリークコミュニティ法のノード集合のサイズが一般に小さいのは，この手法が k-クリークを基にしていて，完全部分グラフという厳しい制約条件がネットワーク全体に一律に適用されてしまうために，ノード集合が細かく分断されやすいからだと考えられる．実世界のデータのネットワーク構造では，必ずしも規則的に完全部分グラフが得られるとは限らないのが普通であり，また本論文で対象としたネットワークのように，ある閾値で絞り込んでいる場合には，わずかな違いでエッジが削除されることもある．そこで，このような場合にも柔軟に対処できることが必要になるが，SR-2 法はその条件を満たしていると思われる．

なお，k-クリークコミュニティ法では，抽出されるノード集合の性質は k に依存する．k によるサイズ分布の違いを，図6に示す．なお，このデータの最大クリークは $k = 8$ である．

この結果から，k を小さくすれば，複雑な形状のコミュニティにフィットできるようになる半面，得られるノード集合のサイズが極端に大きくなってしまい，また k を大きくすると得られるノード集合の数が極端に小さくなってしまう傾向があることがわかる．つまり，k-クリークコミュニティ法では，k の値に極度に依存すると共に，その最適な値を決定するのは難しい．

5 おわりに

本論文では，人間関係のネットワークから活発な人間で構成されるコミュニティ構造を互いの重なりを許容しながら抽出する SR-2 法を提案した．本手法の特性を調べるために，SR-2 法に加えて SR 法と k-クリークコミュニティ法を，実際の Web データから抽出した小規模な人間関係に適用してノード集合を可視化すると共に，抽出性能を評価した．さらに，より大規模なネットワーク論文の共著関係に適用して，抽出ノード集合のサイズ分布から各手法の違いを分析した．この結果として，SR-2 法は，現実の人間の集まりに対応した妥当なコミュニティ構造を抽出できることを示した．

自然界や人間社会にはさまざまなネットワーク構造が存在するが，たとえばある手法があるデータに適していても，別のデータに適しているとは限らない．本論文は，対象とするデータや目的に応じてアルゴリズムに小さな変更を加えるだけで，異なる性質のノード集合を抽出できることを示唆している．この手法の自動的な切り替えには，たとえば単語共起などのネットワークの生成方法やネットワークの統計的な性質などが使えると予想している．

ただし，SR-2 法には，固有ベクトルを量子化する時に，式4の最大値ではなく最初のピークを探索するように変更したことから，抽出されるノード集合が小さくなると同時に抽出回数が増加し，必然的に処理時

間が増加してしまった．今後は，SR 法や SR-2 法を改良して効率を向上させると共に，これらの手法の性質や適合するネットワークの種類をさらに明らかにしていく予定である．

謝辞

日頃より有益な御助言を頂いている上田修功部長，斎藤洋部長，および本実験を手伝って頂いた NTT コムウェアの藤本裕文氏に感謝する．

参 考 文 献

[1] Brin, S. and Page, L.: The anatomy of a large-scale hypertextual Web search engine, in *Proceedings of the 7th International Conference on World Wide Web*, Brisbane, Australia, 1998, pp. 107–117.

[2] Chung, F. R. K.: *Spectral Graph Theory*, CBMS Regional Conference Series in Mathematics, Vol. 92, American Mathematical Society, 1997.

[3] Flake, G., Lawrence, S. and Giles, C. L.: Efficient Identification of Web Communities, in *6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Boston, MA, 2000, pp. 150–160.

[4] Girvan, M. and Newman, M. E. J.: Community Structure in Social and Biological Networks, *the National Academy of Sciences of the United States of America*, 2002, pp. 7821–7826.

[5] Jolliffe, I. T.: *Principal Component Analysis*, Springer Series in Statistics, Springer, 2nd edition, 2002.

[6] 風間一洋, 佐藤進也, 福田健介, 村上健一郎, 川上浩司, 片井修: Web 空間における人間関係を用いた情報探索の一手法, 情報処理学会論文誌：データベース, Vol. 46, No. SIG 13 (TOD27)(2005), pp. 26–39.

[7] Kleinberg, J. M.: Authoritative sources in a hyperlinked environment, *Journal of the ACM*, Vol. 46, No. 5(1999), pp. 604–632.

[8] Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A.: Trawling the Web for emerging cyber-communities, in *Proceeding of the 8th International Conference on World Wide Web*, Toronto, Canada, 1999, pp. 1481–1493.

[9] Manning, C. and Schütze, H.: *Foundations of Statistical Natural Language Processing*, the MIT Press, 1999.

[10] Page, L., Brin, S., Motwani, R. and Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web, Technical report, Stanford Digital Library Technologies Project, 1998.

[11] Palla, G., Derényi, I., Farkas, I. and Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society, *Nature*, Vol. 435, No. 7043(2005), pp. 814–818.

[12] 斉藤和巳, 木村昌弘, 風間一洋, 佐藤進也: ブログ空間の主要トピック抽出, 人工知能学会研究会資料 *SIG-KBS-A501*, 人工知能学会, 2005, pp. 5–10.

[13] Shi, J. and Malik, J.: Normalized Cuts and Image Segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 8(2000), pp. 888–905.

[14] Wasserman, S. and Faust, K.: *Social Network Analysis - Methods and Applications*, Cambridge University Press, 1994.