密な部分構造抽出のための階層的凝集型クラスタリング手法

A Method for Extracting Dense Substructures based on Hierarchical Agglomerative Clustering

高木 允♥　田村 慶一♦　森 康真♦
北上 始♦

Makoto TAKAKI　Keiichi TAMURA
Yasuma MORI　Hajime KITAKAMI

本研究では，Newman らによって提案されているネットワークのクラスタリング手法を改良し，密な部分構造を優先的にクラスタリングする手法を提案する．Newman らの手法は，階層的凝集型クラスタリング手法であり，Q という評価指標をもとに，段階的に大きなクラスタを探索する．この手法は高速であるが，Q の計算式の特徴により，クラスタリングの初期の段階で密な部分構造を壊してしまうという問題がある．この問題により，密な部分構造を見逃してしまう可能性がある．問題点を解決するために，辺の構造情報を辺の重みに換算し、クラスタリングの初期の段階から密な部分構造を優先的に抽出する手法を提案する．

In this paper, we propose a clustering method which clusters the dense substructures preferential by improving the clustering algorithm proposed by Newman et al. The Newman's algorithm is a hierarchical agglomerative clustering method. Newman's algorithm merges clusters based on the value of "Q." This algorithm is fast, however, it may destroy the dense clusters due to the features of the calculating formula of Q. This problem may make us to miss the dense substructures. To address this problem, we propose a method to extract dense substructures preferentially by weighting edges based on the structure information of edges.

1. はじめに

近年，ノードと辺で構成されるネットワークを分類するためのクラスタリング手法が様々研究されている．クラスタリングは，大量の複雑な情報を整理し，有益な知識を発見するための効果的な手法である．ネットワークからのコミュニティ抽出に関する研究は，社会学的な立場やマーケティングの観点からも重要視されている．

大規模なネットワークをクラスタリングするための手法として，Newmanらが高速なクラスタリング手法[1][2]を提案している．文献[1]では，モジュール性Qという評価指標を導入したクラスタリング手法を提案している．Qの値が大きくなるようにクラスタリングを進めるが，クラスタリングを行うたびにQの再計算が必要であり，計算量が増大する．文献[2]では，計算量を削減するために，Qの差分であるΔQのみを計算することでクラスタリングの大幅な高速化に成功している（以降，文献[2]の手法をNewman法と呼ぶ）．

Newman法が評価の対象としているネットワークデータは，重みなし，無向ネットワークである．クラスタとクラスタを併合する際にQの増分であるΔQのみを計算し，ΔQの値が最も高いクラスタを併合していくことで処理の高速化を実現している．Qの値はクラスタとクラスタを併合するときに，併合後のクラスタ内の辺の数が多く，併合後のクラスタからクラスタの外に向かう辺の数が少ないクラスタを併合する場合に高くなる．Newman法では，Q値が最大の時が最もよいクラスタリング結果ということになる．

Newman法ではΔQを計算する際に次数を使用するが，この計算式の特徴により，クラスタリングの初期の段階で次数の低いノード同士優先的に併合されやすい．そのため，密な部分構造をクラスタとして抽出できないという問題がある．ソーシャルネットワークなどからコミュニティを見つけだすような場合，様々な切り口でクラスタを変化させて解析する必要がある．クラスタリング過程を示したデンドログラムをもとに階層を変化させてクラスタリング結果を解析しても密な部分構造がクラスタとして抽出されない．コミュニティとなり得る密な部分構造を発見できなければ，有用なデータを見逃してしまう可能性がある．

この問題点を解決するために，本研究ではNewman法を改良し，クラスタリングの初期の段階で密な部分構造を抽出する手法を提案する．提案手法では，クラスタリングの対象となる重みなし，無向ネットワークの辺に重みを持たせる．全ての辺の重みの初期値を1とし，辺で繋がっている2ノードに共通に繋がっているノード数を2ノード間の辺の重みに足しこむ．さらに，ネットワーク全体の総次数と辺の重みの正規化を行い，ΔQの計算を行っていく．辺に重みを付けることにより，密な部分構造からクラスタリングされていく．

提案手法の有効性を示すために，Newman法と提案手法の比較実験を行った．実験の結果，提案手法は早い段階で優先的に密な部分構造をクラスタリングできており，Newman法では抽出できないクラスタを抽出できる．

本論文の構成は以下のとおりである．2章でNewman法の説明とその問題点について説明し，3章でNewman法の問題点を解決するための提案手法の説明と例を用いた具体的な説明を行う．4章でNewman法と提案手法の比較実験を行った結果を示す．5章で関連研究について述べ，最後の6章でまとめる．

2. Newman 法

本章では，階層的凝集型クラスタリング手法である，Newman 法で用いられる評価値 Q の説明とアルゴリズム，その問題点について説明する．

2.1　評価関数とアルゴリズム

重みなし，無向ネットワーク $G(V,E)$が与えられ，隣接行列 A の要素 A_{uv} が以下のように与えられているとする．ここで，u, v はノード集合 V の要素である．

♥ 学生会員　広島市立大学大学院情報科学研究科博士後期課程　makoto@db.its.hiroshima-cu.ac.jp
♦ 正会員　広島市立大学情報科学研究科　{ktamura, mori, kitakami}@hiroshima-cu.ac.jp

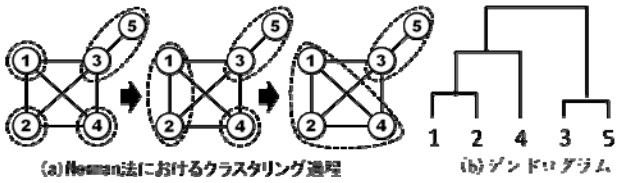

図 1 Newman 法のクラスタリング過程とデンドログラム

Fig.1 Procedure and Dendrogram

$$A_{uv}=\begin{cases}1(\text{ノード}u\text{とノード}v\text{がつながっている場合 }) \\ 0(\text{その他の場合})\end{cases}$$

ノード u の次数 k_u は以下の式で表される．

$$k_u=\sum_{v\in V}A_{uv}$$

ノード u が所属するクラスタ番号を c_u とし，ノード u とノード v とが同じクラスタに所属するかどうかを示す関数を以下のように定義する．

$$\delta(c_u,c_v)=\begin{cases}1(\ c_u\ =\ c_v\text{の場合}) \\ 0(\text{その他の場合})\end{cases}$$

クラスタ内部に存在する辺の割合が多いクラスタリング結果ほど，良いクラスタリングといえる．ここで，クラスタ内部に存在する辺の割合を式で表すと以下の式になる．

$$\frac{\sum_{u,v\in V}A_{uv}\delta(c_u,c_v)}{\sum_{u,v\in V}A_{uv}}=\frac{1}{2m}\sum_{u,v\in V}A_{uv}\delta(c_u,c_v) \quad (1)$$

ここで，m はネットワーク中の辺の総数である．式（1）の値が大きなクラスタリングほど内部に存在する辺の割合が多くなり，評価の高いクラスタリングであるといえる．

しかしながら，この式で表される割合だけでは，ネットワーク全体をひとつのクラスタとするとき最大値1となっているために，評価値としては使用できない．そこで，辺をランダムに張り替えたとき，ノード u とノード v の間に辺が張られる確率 $k_uk_v/2m$ を A_{uv} から引いた値を A_{uv} と置き換えたものをモジュール性の度合い Q として定義する．

$$Q=\frac{1}{2m}\sum_{u,v\in V}\left[A_{uv}-\frac{k_uk_v}{2m}\right]\delta(c_u,c_v)$$

ここで，

$$e_{ij}=\frac{1}{2m}\sum_{u,v\in V}A_{uv}\delta(c_u,i)\delta(c_v,j),$$

$$a_i=\frac{1}{2m}\sum_{u\in V}k_u\delta(c_u,i)$$

とおくと，Q は以下のようになる．

$$Q=\sum_i(e_{ii}-a_i^2)$$

e_{ij} は，ネットワーク全体におけるクラスタ i, j 間に存在する辺の数の割合であり，a_i はネットワーク全体におけるクラスタ i 内に存在する辺の割合を示している．

Newman 法は，Q の値を最大にするようなクラスタリング結果を求める組合せ最適化問題となる．組合せの数はノード数の指数オーダー存在するため，厳密解を求めるのではなく，貪欲アルゴリズムにより，近似最適解を求めている．最初にひとつのノードをひとつのクラスタとし，階層的にクラスタ同士を結合する．どのクラスタ同士を結合するかは，2つのクラスタを結合することにより，Q の値がどれだけ増減するか（$\triangle Q$）で判断する．クラスタ i とクラスタ j とを結合したときの $\triangle Q_{ij}$ は以下の式により求めることができる．

$$\Delta Q_{ij}=2(e_{ij}-a_ia_j) \quad (2)$$

式（2）の評価関数を用いて，$\triangle Q_{ij}$ が最大となる組み合わせを見つけ出し，クラスタリングを続ける．文献[1]においては，クラスタを併合する際に $\triangle Q$ のみを計算しており，$O(nlog^2n)$ という計算量を実現している．ここで，n はネットワーク全体のノード数である．

2.2 Newman 法の問題点

本節では，Newman 法の問題点を説明する．Newman 法は凝集型のクラスタリング手法であるため，アルゴリズムの開始時点ではひとつのノードがひとつのクラスタとみなされている．このとき，式（2）を用いて $\triangle Q$ の値を算出するが，クラスタ i, j 間の辺の数はすべての i, j について 1 であるため，e_{ij} の値はすべてのクラスタ間において同じである．つまり，アルゴリズムの開始時点においては a_i と a_j の値によりどのクラスタとどのクラスタを併合するのかが決定する．次数が低いノードは a の値が低いため，$\triangle Q$ の値が大きくなる．そのため，密に繋がっているノード同士は最初にクラスタリングされにくい．

以下では，問題点を浮き彫りにするために，図 1 に示す 5 ノードから成る小さなネットワークを，Newman 法でクラスタリングし，Newman 法の問題点を示す．アルゴリズム開始時点においては，すべての e_{ij} について e_{ij}=1/14 である．このとき，$a_1 = a_2 = a_4 = 3/14$，$a_3 = 4/14$，$a_5 = 1/14$ である．全ての a_i と a_j の組み合わせにおいて，掛け合わせた時に値が最小になるのは a_3 と a_5 の組み合わせであるため，図 1（a）の左側に示すように，ノード 3 とノード 5 が最初に併合される．さらに処理を続けていくと最終的にノード 1，2，4 から成るクラスタとノード 3, 5 から成るクラスタが抽出される．

しかしながら，ノード 3 からはノード 1，2，4 で構成されたクラスタと 3 本の辺で繋がっており，ノード 1，2，3，4 から構成されるクラスタが抽出されるのが自然である．図 1（b）に示すデンドログラムからも，どの階層でクラスタリング結果を解析してもノード 1，2，3，4 から構成されるクラスタは抽出できないことが確認できる．

ネットワークの規模が大きな場合は，クラスタリングの初期の段階で密な部分構造が分割してクラスタリングされても，クラスタリングが進むと，クラスタ同士が併合して，分割された密な部分構造がひとつの大きなクラスタを形成することはある．しかしながら，例で示したように初期の段階で密な部分構造を分割してしまう可能性が高い．このような分割が起こると，デンドログラムをもとにクラスタリング結果を変化させてクラスタの解析を行っても密な部分構造を見つけだすことはできない．評価実験で大規模なネットワークに Newman 法を適用したときの問題点を示す．

3. 提案手法

本章では，2.2 節で述べた Newman 法の問題点を解決するための手法を提案する．

3.1 アルゴリズム

提案手法の基本的な考えは，辺のつながりをもとに，辺に重みを付けることである．ネットワーク中のノードを u, v で表現し，u, v 間の辺の重みを w_{uv} とする．また，ノード u に隣接しているノードの集合を $Adj(u)$ と表現する．以下に提案手法のアルゴリズムを示す．

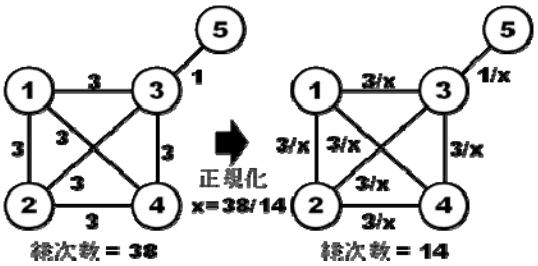

図 2 辺への重み付けと正規化

Fig. 2 Weighted Edges and Regulation

ステップ 1 辺への重み付け

辺の重みの初期値として，全ての w_{uv} について w_{uv}=1 とする．次に，全てのノード u について，$Adj(u)$を求める．さらに，全てのノードのペア u, v について積集合 $Adj(u) \cap Adj(v)$ を求め，ノード u とノード v ともに繋がっているノードの数 $|Adj(u) \cap Adj(v)|$ を w_{uv} に足しこむ．つまり，辺で繋がっているノードに共通に繋がっているノードの数を辺の重みに足しこんでいく．この操作を行うことで，密な部分構造が浮き彫りになるため，Newman 法の式（2）を用いた場合，密な部分構造からクラスタリングされやすくなる．ここで，$Adj(u) \cap Adj(v)$をノード u, v の共通ノード，$|Adj(u) \cap Adj(v)|$ をノード u, v の共通ノード数という．

ステップ 2 辺の重みの正規化

辺に重みを付けた場合，Newman 法で使用している式（2）を用いると総次数（分母）が変化してしまう．そのため，総次数を，重みを付けない場合の値に戻す必要がある．そのために，総次数と辺の重みの正規化を行う．

ステップ 3 Newman 法の実行

正規化した重みをもとに Newman 法の評価関数である式（2）を用いてクラスタリングしていく．辺への重み付け以外の処理は，Newman 法のアルゴリズムを使用しているため，辺の総数を m とすると，ステップ 1 での計算量は $O(m)$ となり，全体の計算量は $O(m)+O(nlog^2n)$となる．

以上のステップを経て，共通ノード数をもとに辺への重み付けを行うことで，クラスタリングの初期の段階においても密に繋がっているクラスタ間の e_{ij} の値が大きくなり，密に繋がっているクラスタ同士が優先的にクラスタリングされやすくなる．また，計算量の大幅な増加はないため，処理時間の大幅な増大はないと考えられる．

3.2 提案手法の例

本節では，具体的な例を用いて提案手法の説明を行う．図 2 に示す 5 ノードからなるネットワークへの重み付けとクラスタリングを考える．

ステップ 1

全ての辺の重みを 1 とする．ノード 1, 2 の共通ノードは，$Adj(1) \cap Adj(2) = \{3, 4\}$であり，共通ノードの数である $|Adj(1) \cap Adj(2)|$ は 2 となる．辺で繋がっている全てのノード間について共通ノード数の算出を行い，辺の重みの初期値である 1 に共通ノード数を足すことで，図 2 左側の結果が得られる．ノード 3 とノード 5 では，$Adj(3) \cap Adj(5) = \phi$ となり，共通ノードが存在しないため，辺の重みは 1 のままである．

ステップ 2

次に，総次数の変化を防ぐために，正規化を行う．図 2 の例では，重みを付けない場合の総次数は 14 であるが，重みを付けると 38 に変化する．そこで，総次数を 38 から 14 に戻し，w_{uv} を(38/14)で割ったものを辺の重みとする．

ステップ 3

上述した 2 つのステップで得られた重みをもとに，式（2）を用いてクラスタの併合を繰り返していく．図 2 の例をクラスタリングすると，最初にノード 3，5 を併合したときの Q の値が Newman 法の場合と比べて減少し，ノード 1，4 が最初にクラスタリングされる．ノード 5 は最後にノード 1, 2, 3, 4 から成るクラスタに併合されてクラスタリングを終了する．よって提案手法を用いるとノード 1，2，3，4 から成る完全グラフがクラスタとして抽出可能となる．

表 1 実験に用いたネットワークデータ

Table 1 Network Data for Evaluations

ネットワーク	ノード数	辺数	密度
ブログ	80	133	0.042089
ワード	7,207	31,784	0.001224

4. 性能評価

性能評価として，提案手法と Newman 法の比較実験を行った．Q 値の比較，階層構造の比較，クラスタの密度の比較，クラスタリング結果の解析の比較の 4 つの観点から行った比較実験について示す．評価実験には，文献[9][10]で得られた頻出なブログネットワークと文献[11]で提供されているワードネットワークの一部を用いた．各ネットワークデータの詳細を表 1 に示す．表 1 中の密度は，$n/\{m(m\text{-}1)/2\}$で算出した[12]．n はネットワーク全体のノード数であり，m はネットワーク全体の辺の数である．今回は，小規模と大規模なネットワークとして，ノード間の繋がりを把握しやすく，結果として得られるクラスタが妥当かどうか判断しやすいため，上記 2 つのネットワークを選んだ．

4.1 Q 値の比較

各手法において，ノードの併合が進むにつれ，どのように Q 値が変化していくのかの比較を行った．提案手法においては，辺に重みを付けていない状態の Q 値を算出するために，提案手法におけるノードの併合情報（併合の順番を保持したリスト）を用いて，Newman 法により各ステップにおける Q 値の再計算を行った結果を示している．図 3 にブログネットワークでの実験結果，図 4 にワードネットワークでの実験結果を示す．グラフの横軸は x 回目のクラスタの併合を示しており，縦軸は Q 値を示している．

ブログネットワークでは，Q の最大値は Newman 法の方が高くなっている．Newman 法の Q の最大値は 0.6084 で，提案手法の Q の最大値は 0.5802 であった．しかしながら，提案手法では 50 回目あたりのクラスタの併合までは Newman 法よりもよい結果となっている．

提案手法はアルゴリズムの初期段階で密な部分構造を優先的にクラスタリングしていくためである．つまり，アルゴリズムの初期の段階でクラスタ内の辺の数が多く，クラスタ間の辺の数が少ないクラスタが形成されていることになる．

ワードネットワークでは，Q の最大値，クラスタリングの途中段階における Q 値ともに提案手法が上回っている．Newman 法の Q の最大値は 0.4640 で，提案手法の Q の最大値は 0.5256 であった．この原因としては，ワードネットワークは比較的疎なネットワークデータであるということが挙げられる．疎なネットワークの場合，提案手法を用いると密なネットワークと比べ，ネットワーク中の密な部分構造をより識別しやすいためである．

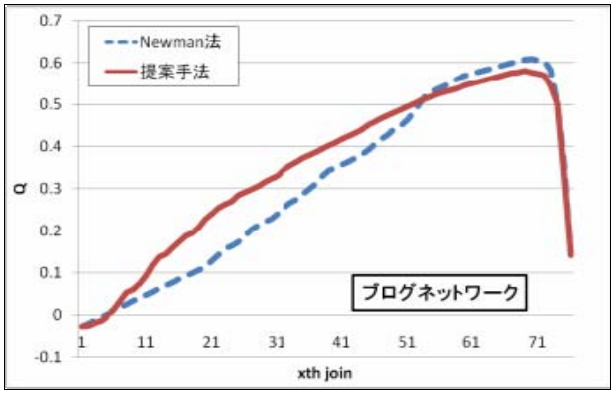

図 3　ブログネットワークでの Q 値の比較

Fig. 3　Comparison of Value of Q (blog network)

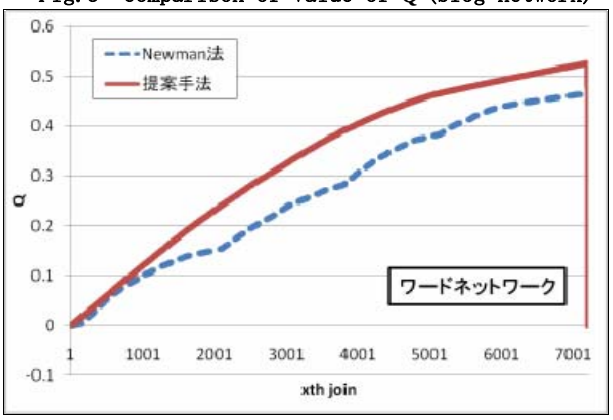

図 4　ワードネットワークでの Q 値の比較

Fig. 4　Comparison of Value of Q (word network)

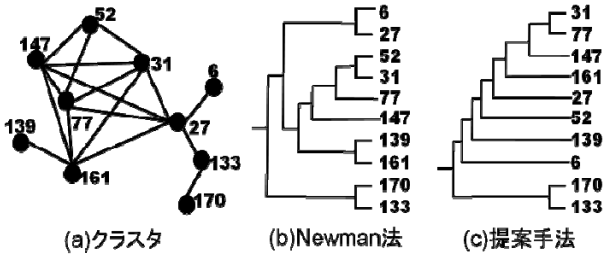

図 5　ブログネットワークのデンドログラムの比較

Fig. 5　Comparison of Dendrogram (blog network)

4.2　階層構造の比較

本節では，両手法によるクラスタリング過程を調べるため，ブログネットワークのクラスタリング過程を示した階層構造を可視化し，解析を行った．図 5 (a) は両手法において Q 値が最大の時点で抽出されたクラスタのひとつである．ここで，提案手法における Q の最大値は，提案手法におけるクラスタの併合情報を用いて式 (2) で Q の再計算をした値のことである．Q 値が最大の時点では両手法において図 5 (a) に示すクラスタを構成するノードは全て同じであった．

このクラスタのクラスタリング過程を示すデンドログラムをそれぞれ図 5 (b) と図 5 (c) に示す．まず，Newman 法におけるクラスタリング過程について説明する．図 5 (b) から，ノード 6 とノード 27，ノード 139 とノード 161 が初期の段階でクラスタリングされていることが分かる．3.2 節でも示したように，次数が低いノードが初期の段階で併合されやすいことを示している．ノード 27，31，77，147，161 は 5 つのノードから成る完全グラフを形成しており，クラスタリングの途中段階でひとつのクラスタとして抽出されることが望まれる．しかしながら，Neman 法ではノード 6 とノード 27 が初期の段階でクラスタリングされているため，クラスタリングのどの階層においても 5 ノードから成る完全グラフをクラスタとして抽出できない．

次に，提案手法におけるクラスタリング過程について説明する．図 5 (c) から，Newman 法とは異なり，クラスタリングの初期の段階ではノード 31 とノード 77 がクラスタリングされ，階層を上げると先ほど説明した 5 つのノードから成る完全グラフを形成していることがわかる．その後，次数の低いノードをクラスタリングしていき，最終的に Newman 法と同じクラスタリング結果を得ている．デンドログラムの比較から，提案手法は密な部分構造を優先的にクラスタリングしていることが確認できた．

この例では Q が最大となる階層のクラスタリング結果を示している．しかしながら，様々な角度からネットワークを解析したい場合などは階層を上下させてクラスタリング結果を解析する必要がある．コミュニティのような密な部分構造を崩さずに解析を行いたい場合などは，提案手法の性能を発揮することができる．

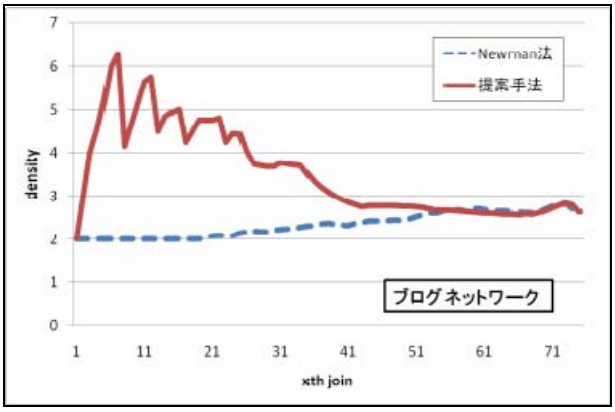

図 6　ブログネットワークでの密度の比較

Fig. 6　Comparison of Density (blog network)

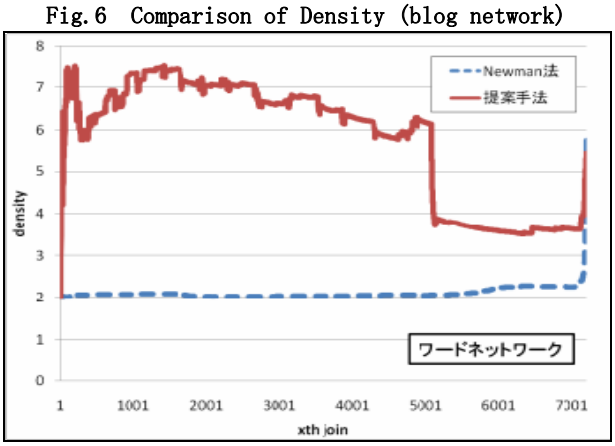

図 7　ワードネットワークでの密度の比較

Fig. 7　Comparison of Density (word network)

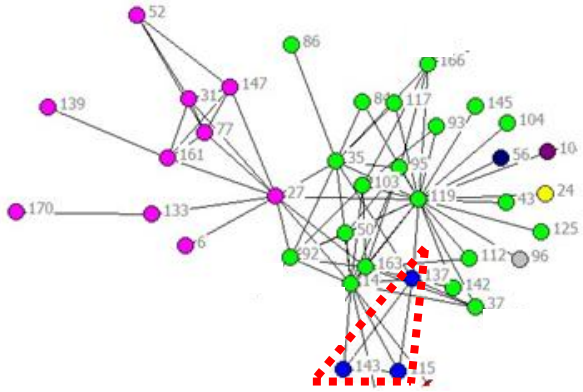

図 8　提案手法における 60 ステップ目のクラスタリング結果の一部

Fig.8　A Part of Clustering Result of 60th Step in Proposed Method

4.3　クラスタの密度の比較

次に，両手法のクラスタリング過程において形成されるクラスタ密度の比較を示す．密度の評価としては，各ステップにおいて形成されたクラスタ内のノードの数と辺の数を用い，それにクラスタサイズを掛け合わせることで，クラスタ密度の評価を行う．つまり，密度が高く，クラスタサイズが大きい場合に評価値が高くなる．クラスタ i の密度は以下の式を用いて算出した．

$$density_i = \frac{m_i n_i}{n_i(n_i-1)/2} = \frac{m_i}{(n_i-1)/2} \qquad (3)$$

ここで，n_i はクラスタ i 内のノード数，m_i はクラスタ i 内で繋がっている辺の数である（クラスタ i から外に向かっている辺は数えない）．2 つのノードから成るようなサイズの小さなクラスタは密度が高くなってしまうため，サイズ n_i を掛け合わせて，サイズの大きなクラスタの評価値を上げる．各ステップにおいてノード数 1 のクラスタは無視し，サイズが 2 以上のクラスタの密度を対象としている．

各ステップにおいて計算したクラスタの密度の平均をそれぞれ図 6 と図 7 に示す．横軸は x 回目の併合を示しており，縦軸は式 (3) で求めた密度を示している．図 6，図 7 より，提案手法における密度が Newman 法に比べて高いことが分かる．最終的にはネットワーク全体がひとつのクラスタになるため，クラスタリングが終了する段階では，密度は同じ値に収束する．図 6 と図 7 から，提案手法ではクラスタサイズが大きく密度が高いクラスタがアルゴリズム開始時点から形成されていることが分かる．

デンドログラムを用いた階層構造の解析と，密度の比較において，提案手法では Newman 法に比べより密な部分構造を優先的にクラスタリングしていることが分かった．

4.4　ブログネットワークにおけるクラスタの評価

本節では，ブログネットワークを両手法においてクラスタリングした結果，デンドログラムのある階層においてどのような内容のクラスタが形成されているのかを調べた結果の一部を示す．文献[9][10]で使用しているブログネットワークは，ブロガーをノード，トラックバックによるつながりを辺としたネットワークである．

図 8 は提案手法における 60 ステップ目のクラスタリング結果の一部を示している．図中の点線で囲まれている 3 ノード（ノード番号 115, 137, 143）から成るクラスタの話題は，

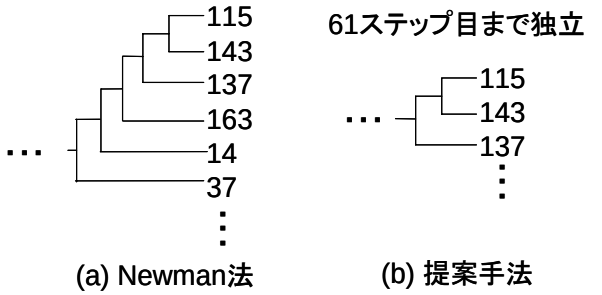

図 9　クラスタの形成と併合

Fig.9　Constitution and Merge of Clusters

tf-idf 法や手作業による調査で，日本のプロ野球チームである読売ジャイアンツのファンのブロガー集団であることが分かった．この 3 つのノードから成るクラスタに繋がっている他のノードは，主にその日にあったスポーツの結果など，スポーツ全般の話題について言及しているブロガーであった．

図 9 に 3 つのノードから成るクラスタの形成と併合を示したデンドログラムを示す．図 9 から，このクラスタは Newman 法と提案手法の両手法においてデンドログラムの高さ 2 の段階で形成されている．図 9 (a) に示すように，Newman 法では，デンドログラムの高さ 3 の段階でこのクラスタは他のクラスタと併合していた．しかしながら，提案手法では図 9 (b) に示すように，デンドログラムの高さ 61 までこのクラスタは独立して形成されていた．この事実から，提案手法では，点線で囲まれた 3 ノードから成るクラスタは他のクラスタと独立性が高く，このクラスタに繋がっている他のノード同士のつながりが強いとみなされていることが分かる．

Newman 法ではデンドログラムの高さ 3 の段階で他のクラスタと併合している．これは，このクラスタと繋がっている周りのクラスタ同士が密に繋がっているにも関わらず，その密なつながりを考慮していないために起こった現象である．このことから，辺のつながりによる重みを考慮する提案手法では，コミュニティとなりうるクラスタを発見しやすくなる．

5. 関連研究

近年，タンパク質の相互作用ネットワーク，ソーシャルネットワークやブログネットワークなどをクラスタリングし，コミュニティや意味のある部分構造を見つけ出す研究に Newman法が適用されている[3][4][5]．また，Newman法を改良することにより，クラスタリングの精度を向上させる手法[6]や，より大規模なネットワークを高速にクラスタリングするための手法[7]，Newman法を，重複を許したクラスタリング手法に改良する研究[8]などが提案されている．文献[9]では，Newman法の改良を行い，コアなコミュニティとコミュニティ間の橋渡しをしているノードを見つけ出すことに成功している．この章では，本研究の関連研究として，辺に重みを持たせている文献[6]とクラスタとクラスタの併合を効率的に行うことを考えている文献[7]に注目する．

文献[6]では，ブログの記事をノード，トラックバックによるつながりを辺としたブログネットワークにNewman法を使用すると，多数の小さなサイズのクラスタと少数の大きなサイズのクラスタが抽出され，大きなサイズのクラスタでは複数の話題が含まれていたという報告がされている．文献[6]ではノードであるブログ記事の類似性に基づき，辺に重みを

持たせることにより，この問題点を解決している．

提案手法では，ノードの類似性は考慮せず，辺の構造のみを解析して重み付けを行っている点で文献[6]とは異なる．ノードの類似性の計算を行わず，辺のつながりの情報のみで重み付けを行うため，様々なネットワークに適用可能な汎用性を持たせた手法であるという点で優れている．また，本研究では階層構造の情報を利用してクラスタの解析を行うことを前提としており，ある特定のクラスタリング結果を最終結果としている文献[6]とは異なる．

文献[7]では，Newman法を改良し，クラスタを効率よく併合させることでNewman法の高速化を実現している．実験結果から，Newman法では扱えない規模のネットワークをクラスタリングできたと報告している．この手法は，Q値を最大化するようにクラスタを合併していないが，最終的に得られた結果はNewman法よりもQ値が高い結果となっており，処理時間とQ値の面で高い性能を示している．

提案手法では，辺に重みを付けることにより，密な部分構造を優先的に確保していくことを目的としている．文献[7]においては，クラスタを平均的に成長させていくことで処理の効率化を図っている．クラスタ同士の合併効率の優先順位を高くしているため，本研究が目的としている密な部分構造の優先的なクラスタリングは困難となり，密な部分構造が複数のクラスタに分割される可能性がある．本研究では，密な部分構造が複数のクラスタに分割されないという点で優れている．今後は，極大クリーク列挙アルゴリズムとの比較をする予定である．

6. おわりに

本研究では，階層的凝集型クラスタリング手法である Newman 法を改良した．提案手法は，辺の構造を解析し，辺に重みを付けることにより，密な部分構造を優先的にクラスタリングしていく手法である．Newman 法と提案手法の比較実験を行った結果，提案手法は Newman 法と比べ，密な部分構造を優先的にクラスタリングできていた．ブログネットワークにおける評価実験で，デンドログラムをもとにクラスタリング結果を解析すると，コミュニティとなりうるクラスタを識別できていたことが分かり，提案手法の有効性を示すことができた．

今後の課題としては，様々な種類のネットワークでの評価実験や，クラスタリング過程や結果をより理解しやすくするためのユーザインタフェースの開発が挙げられる．

[謝辞]

本研究の一部は，日本学術振興会・特別研究員奨励費（課題番号：18・0205），日本学術振興会・科学研究費補助金（基盤研究（C）（一般），課題番号：20500137）の支援により行われた．

[文献]

[1] M. E. J. Newman and M. Girvan: Finding and Evaluating Community Structure in Networks. Physical Review E 69, 026113 (2004).

[2] Clauset, A., Newman, M.E.J. and Moore, C.: Finding Community Structure in Very Large Networks, Physical Review E, Vol.70, p.066111 (2004).

[3] 高木 允，森 康真，田村 慶一，北上 始: ブログユーザ空間からの重複を許した頻出コミュニティ抽出法，情報処理学会論文誌，数理モデル化と応用，Vol.49，No. SIG 4 (TOM20) pp.93-104, 2008 年 3 月．

[4] M. Takaki, Y. Mori, K. Tamura, S. Kuroki and H. Kitakami :Method for Extracting Frequent Communities from Blog User Spaces, The 2007 International Conference on Parallel & Distributed Processing Techniques & Applications (PDPTA'07), Vol. II, pp.773-779, July 2007.

[5] D. L. Nelson, C. L. Mcevoy, and T. A. Schreiber: The University of South Florida Word Association, Rhyme, and Word Fragment Norms.

[6] 安田雪: ネットワーク分析―何が行為を決定するか，新曜社，1997.

[7] Andre X. C. N. Valente and Michael E.Cusick: Yeast Protein Interactome Topology Provides Framework for coordinated-functionality, Nucleic Acids Research, Vol.34, pp. 2812, 2006.

[8] 湯田 聴夫，小野 直亮，藤原 義久: ソーシャル・ネットワーキング・サービスにおける人的ネットワークの構造，情報処理学会論文誌，Vol.47, No.3, pp.865-874(2006).

[9] Noor F. Ali-Hasan and Lada A. Adamic: Expressing Social Relationships on the Blog through Links and Comments, International Conference on Weblogs and Social Media (ICWSM), 2007.

[10] 安藤 潤，吉井 伸一郎: WWWナビゲーション向けコミュニティ分割手法に関する一考察，情報処理学会研究報告知能と複雑系，pp.115-122(2006).

[11] K. Wakita and T. Tsurumi: Finding Community Structure in Mega-Scale Social Networks, Proceedings of the 16th International Conference on World Wide Web 2007, pp.1275-1276.

[12] M. Takaki, K. Tamura, Y. Mori, and H. Kitakami: A Extraction Method of Overlapping Cluster based on Network Structure Analysis, the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, Workshops on IWI2007, the IEEE Computer Society Press, pp.212-216, 2007.

高木 允 Makoto TAKAKI
広島市立大学大学院情報科学研究科博士後期課程在学中．2006 同大学院情報科学研究科博士課程前期修了．日本データベース学会，情報処理学会 各学生会員．

田村 慶一 Keiichi TAMURA
広島市立大学大学院情報科学研究科講師．2000 九州大学大学院システム情報科学研究科博士前期課程修了．博士（情報科学）．情報処理学会，日本データベース学会， IEEE CS 各会員．

森 康真 Yasuma MORI
広島市立大学大学院情報科学研究科助教．1994 北陸先端科学技術大学院大学情報科学研究科博士前期課程修了．情報処理学会，電子情報通信学会，人工知能学会，IEEE CS，ACM 各会員．

北上 始 Hajime KITAKAMI
広島市立大学大学院情報科学研究科教授．1976 東北大学大学院工学研究科博士前期課程修了．博士（工学）．データベースおよび人工知能などの研究開発に従事．日本データベース学会論文誌編集委員，人工知能学会評議員，情報処理学会（TOM）論文誌編集委員，DEWS2008 組織委員，情報処理学会，電情報通信学会，IEEE および ACM 各会員．