大規模２部グラフの可読性向上のためのクラスタ構造の動的描画

Interactive Drawing Techniques for Clustered Structures in Bipartite Graphs

佐藤 修治　　三末 和男　　田中 二郎*

Summary.　ネットワークの構造を理解するためには，可視化を行うことが有効である．近年の情報量の増大に伴い，大規模なグラフに対応した描画手法の開発が必要である．本論文ではまず，対象とするグラフのノードをクラスタリング手法を利用して分類する手法を提案する．そして，クラスタの生成を動的に表示する「クラスタリングスライダ」と，クラスタを等高線のように表示する「等類似度線描画」という２つのインタラクティブな可視化手法を用いて，クラスタリングの結果を効果的に表示することで可読性の向上を狙う．最後に描画結果から新たな知見を述べ，提案手法の有効性を示す．

1　はじめに

本研究の目的は，大規模グラフの可視化における「わかりやすさ」の向上である．

情報インフラの整備の浸透に伴い，我々が手に入れられる情報の量は爆発的に増大している．これら情報を人間が知識として得る際には，その情報を抽出し「可視化」を行うことで人間の理解を支援することが可能となる．そのため，可視化技術は様々な分野で研究対象とされている．しかし，一般的な手法ではデータが大規模になるほど可視化結果の可読性は低くなるため，可視化手法を検討する必要がある

本論文では大規模なグラフを可読性の維持された表現での可視化を行う．ノードをクラスタリング手法を利用して分類し，先行研究である「アンカーマップ [1]」の描画手法に適用する．これにより表現される大規模ネットワークを，クラスタの情報をインタラクティブに表現することで可読性の向上を狙う．

2　前提知識

2.1　２部グラフ

グラフとは，ノードの集合とその接続関係を表すエッジの集合から構成される論理構造であり，物の間の関係を表すのに適している．グラフの内，ノードの集合を二つの排他的な集合 V_1 と V_2 に分割することができ，エッジの集合 E が $V_1 \times V_2$ の部分集合であるようなグラフを２部グラフといい，$G = (V_1 \cup V_2, E)$ として表現される．２部グラフは実世界の様々な場面で現れ,「顧客」と「購買商品」の関係,「コミュニティ」と「メンバー」の関係などが例に挙げられる．本論文では２部グラフとして表現されるネットワークに焦点をあてる．

* Shuji SATO, Kazuo MISUE and Jiro TANAKA, 筑波大学大学院 システム情報工学研究科コンピュータサイエンス専攻

2.2　アンカーマップ

グラフの可視化に良く用いられる描画法の一つとして Eades のスプリングモデル [2] がある．スプリングモデルは，エッジにバネを埋め込み安定状態を計算することで，ノードの配置を求めるグラフの描画手法である．

アンカーマップはスプリングモデルを発展させ，２部グラフの２種類のノードの集合の一方に位置の制約を課した描画スタイルである．本研究での位置の制約は次の２点である．

- 集合 V_1 の要素を円周上に等間隔に配置される
- 集合 V_2 の要素は自由配置で，集合 V_1 の要素との関係を適切に表現する位置に配置される

このとき，集合 V_1 を「アンカーノード」と呼び，集合 V_2 を「フリーノード」と呼ぶ．エッジはアンカーノードとフリーノードを直線で接続する形式で表現される．

図１にアンカーマップのスタイルで可視化した結果を示す．データはある施設での商品の購買履歴を元にしている．アンカーノードは商品が購入された時間帯で,「00h」～「23h」の 24 個のノードが存在する．描画時にはノード「00h」を最北端に位置し，時計周りで 24 個のノードを配置した．フリーノードは購入された 37 種類の商品である．エッジは商品とそれが購入された時間に接続されており，スプリングの強さは購入された商品数に比例して強く設定されている．アンカーマップのスタイルでは，ノードの配置から「午前中は買い物が少なく，夕方から夜にかけて頻繁に買い物がなされている」といったことを直観的に読み取ることが可能である．また，ラベルを読み取ると,「正午近くと 18 時から深夜にかけて食品が売れ，飲料製品は時間帯に関係なく購入されている」ということも把握することが出来る．スプリングモデルと比較して着目するべき概念の関連性が明確になり，ネットワークの構造的特徴を認

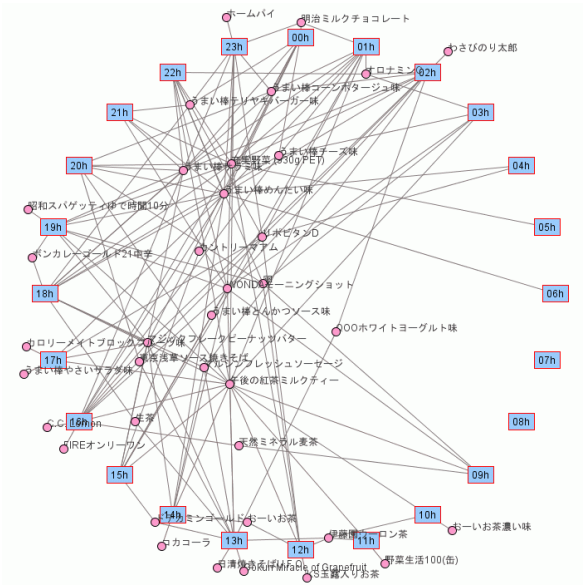


図 1. アンカーマップ

知するのに効果的である．

2.3 大規模グラフ描画における問題

既存手法のアンカーマップを用いて大規模グラフを描画した際，可読性において問題が生じる．大規模なネットワークにおいては形成するノードやエッジが多く存在する．表示するデバイスの解像度を上げることできめ細かい描画を行い，すべてのノードとエッジを表示することは可能であるが，人間の理解度の向上には繋がりにくい．人間がネットワークを認識・理解するのに適当な大きさの領域に，簡潔に描画することが重要となる．

従来の描画アルゴリズムは人間の理解度を上げるために，エッジの交差数や総線長を最小にする等のグラフのレイアウトを変更する手法を用いており，そのレイアウトの指標にに美的基準を採用してきた [1, 3]．ここで，可視化対象が大規模になりノード数が増加した場合，ノードのレイアウトの変更だけで可読性を向上させることは難しい．

3 可読性向上へのアプローチ

大規模なネットワークの可読性を向上させるため，関連性の強いいくつかのノードやエッジをまとめるクラスタリング処理を行うことにした．

ネットワークの可視化の目的は，読み手や場面により異なる．人間の認識能力は各々の読み手で異なるため，画面上に配置されているノードの数がどの程度で適切であると感じるかは人によって異なる．加えて，その時に注視したいものも目的により異なる．可読性を向上させるためには，時と場合で適切な規模の描画結果を出力することが必要である．

クラスタリングは，ノードの集合を分類すると共にその関係の類似性の情報を付加するものである．クラスタリングの結果を有効に可視化する 2 つの手法を提案し，実装方法について述べる．

3.1 クラスタリング

クラスタリングとは、異なる性質のもの同士が混在している集合の中から互いに類似したものを集めてクラスタを形成することで，集合を分類する手法である [4]．

一般的にグラフ描画におけるクラスタリングは，論理的に同じ接続構造を持つノードを一塊にして考えるものであるが，可視化する対象が大規模になればなるほど，同じ接続構造をもつノードの割合は減少すると考えられるので，これらのノードのみをクラスタ化しても効果は薄い．本稿で扱うクラスタリング手法は，ノード間のエッジの接続関係から類似度を計算し，階層構造のクラスタを生成するものである．ただし，クラスタが 2 部グラフの 2 種類の集合を跨いで形成されることはないようにして，2 部グラフの構造を保持するよう注意する．以下ではフリーノードのクラスタリングについて述べるが，対称性を利用してアンカーノードにも同様の処理をすることが可能である．

ノード間類似度の求め方

ノード x が接続するノードの集合を $A(x)$ とする．ノード x とノード y の類似度 $S(x,y)$ を次式で表す．

$$S(x,y) = \frac{|A(x) \cap A(y)|}{|A(x) \cup A(y)|} \quad x, y \in \mathbf{V} \qquad (1)$$

例として，$A(x) = \{\alpha, \beta, \gamma\}$，$A(y) = \{\alpha, \beta\}$ としたとき，$S(x,y)$ は以下のように算出される．

$$\begin{aligned} S(x,y) &= \frac{|\{\alpha, \beta\}|}{|\{\alpha, \beta, \gamma\}|} \\ &= \frac{2}{3} \approx 0.66 \end{aligned}$$

これをフリーノード V_2 のすべての組み合わせにおいて計算する．結果は，三角行列の表で表すことが出来る．

クラスタの構築

全ノードの組み合わせの類似度を求めた後，類似度の高い順にクラスタを形成していく．

クラスタ $\mathbf{C}$ はノードとクラスタの集合であり，以下の式で定義する．

$$\begin{aligned} \mathbf{C} &= \{c_1, \cdots, c_n\} \quad (n \geq 2) \\ c_1, \cdots, c_n &\in \mathbf{V} \cup \mathbf{C} \end{aligned}$$

クラスタにおける類似度 $S(C)$ は以下のように定義する．

$$S(\mathbf{C}) = \max_{1 \le i < j \le n} S(c_i, c_j) \quad (c_i, c_j \in \mathbf{V} \cup \mathbf{C})$$

式 (1) に加え，類似度 $S(p,q)$ $(p,q \in \mathbf{V} \cup \mathbf{C})$ を以下のように定義する．

$S(p,q) = S(q,p)$　可換性

$$S(p,q) = \begin{cases} \dfrac{|A(p) \cap A(q)|}{|A(p) \cup A(q)|} & (p,q \in \mathbf{V}) \\ \max\limits_{\forall q' \in q} S(p,q') & (p \in \mathbf{V}, q \in \mathbf{C}) \\ \max\limits_{\forall p' \in p, \forall q' \in q} S(p',q') & (p,q \in \mathbf{C}) \end{cases}$$

類似度が上記の定義になるようなクラスタを構築する手順を以下に示す．

1. $s = \max\limits_{\forall x, \forall y \in V_2, x \neq y} S(x,y)$ となる s を求める．
2. $S(x,y) = s$ となるノード x，y が存在する場合，$x \in C_x$，$y \in C_y$ を満たすクラスタ C_x，C_y が存在するかを確認する．
   (a) C_x と C_y が両方存在しない場合，x と y を含み，類似度 $S(C) = s$ のクラスタ C を生成する．
   (b) C_x が存在し，C_y が存在しない場合，$C = G(C_x)$ とを満たすクラスタ C を求める．
      i. $S(C) = s$ の場合，C に y を加える
      ii. $S(C) > s$ の場合，y と c を含み，類似度 $S(C') = s$ のクラスタ C' を生成する．
   (c) C_y が存在し，C_x が存在しない場合，2b の x と y を入れ替えて同様の動作を行う．
   (d) C_x と C_y が両方存在する場合，$C'_x = G(C_x C_y)$ と $C'_y = G(C_y)$ を求める．
      i. $s = S(C'_x) = S(C'_y)$ の場合，C'_x と C'_y を結合する
      ii. $s = S(C'_x)$ かつ $s < S(C'_y)$ の場合，C'_x に C'_y を加える．
      iii. $s = S(C'_y)$ かつ $s < S(C'_x)$ の場合，C'_y に C'_x を加える．
      iv. それ以外の場合は，C'_y と C'_x を含み，類似度 $S(C) = s$ のクラスタ C を生成する．
3. s であるノードの組み合わせが残っていれば，2 に戻る
4. $s > 0$ の場合，s より小さい類似度の値中次に大きいものを s として，2 に戻る

	v_1	v_2	v_3	v_4
v_1	-	0.9	0.8	0.72
v_2		-	0.9	0.81
v_3			-	0.9
v_4				-

表 1. 類似度表

G(C) は C の最上位の親クラスタで，以下の式で定義する．

$$\begin{aligned} &\text{When} \quad \mathbf{C}' \supseteq C' \supseteq C \\ &G(C) = C' \\ &\text{satisfied} \quad S(C') = \min_{\forall c' \in \mathbf{C}'} S(c') \end{aligned}$$

以上を行うことにより，2 つのノードの組み合わせの結果から全ノードのクラスタを木構造で構築していくことが出来る．エッジについては，それぞれの接続ノードを包含するクラスタにも同様に接続する．

本手法の特徴

このクラスタリング手法の特徴として，通常の階層型クラスタリングよりも計算速度が速いことが挙げられる [5]．通常の階層型クラスタリング手法は，

1. すべてのノードの組み合わせの内最も高い類似度の要素を 1 組探索し，クラスタとして結合させる
2. 残りのノードの集合と生成されたクラスタの類似度を再計算

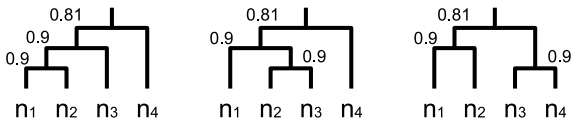
を繰り返し，階層構造を形成していく．それと比較し，本手法は類似度の計算は 1 回で良いので，類似度計算自体のオーダは $O(n^2)$ にとどめることが出来る．

もう一つの特徴は，ノードの並び順（クラスタの結合順序）によりクラスタリング結果が変わらない一貫性である．例として，4 つのノード $v_1, \cdots, v_4$ の接続ノードの集合 $A(v_1), \cdots, A(v_4)$ を

$$\begin{aligned} A(v_1) &= \{2, 3, 4, \cdots, 10\} \\ A(v_2) &= \{1, 2, 3, \cdots, 10\} \\ A(v_3) &= \{1, 2, 3, \cdots, 9\} \\ A(v_4) &= \{0, 1, 2, \cdots, 9\} \end{aligned}$$

と定義する．類似度を計算した結果を表 1 に示す．

ここで，類似度が 0.9 のものが 3 つあるが，結合する順番はノードの並び順に依る．すなわち，v_1-v_2，v_2-v_3，v_3-v_4 のどれを始めにクラスタとするかである．それぞれを最初にクラスタリングを行い，全ノードをクラスタリングした場合の結果をデンドログラム（樹状図）として表したのが，図 3.1 である．v_1-v_2 と v_2-v_3 は論理的に同じ構造であるが，v_3-v_4 は他二つとは異なる構造になる．



(a) v_1–v_2 (b) v_2–v_3 (c) v_3–v_4

図 2. クラスタリング順による変化

本論文のクラスタリング手法でこれらノード集合のクラスタリングを行うと，どの順番で行っても，v_1〜v_4 は類似度 0.9 のクラスタで結合され，結果の一貫性が保たれる．通常のクラスタリングよりも高い類似度で結合されるクラスタが多くはなるが，可読性は変わらないと考える．

図 3(a) は閾値を $t = 0.50$ 図 3(b) は $t = 0.35$ でクラスタリングを施したアンカーマップ描画である．混み具合がかなり解消されており，クラスタリングの効果により可読性が向上したことを確認することが出来る．

3.2 クラスタリングスライダ

グラフ全体の規模を読み手にとって最適にし，クラスタの性質を認識する点において有効な手法として，クラスタリングの表示レベルをインタラクティブに変更することが可能な「クラスタリングスライダ」を開発した．スライダはポインティングデバイスにより操作でき，自由にクラスタリングの強さを変更できる．

スライダの値は最大値 101[%]，最小値 0[%] を設定している．ノード間類似度の最大値は 100% であるため，スライダの値が 101% に位置しているときには，クラスタ化されていないオリジナルのノードの集合を描画する．また逆に最小値に位置しているときは，全ノードが一つのクラスタにまとめられ，画面上には 1 つのクラスタのみを表示する．

実装方法

スライダの位置している値の類似度以上のクラスタを表示するというのは，デンドログラムを水平に切断し，切断面のエッジに接続するノードとクラスタのみを表示するのと同義である．スライダを動かした際にクラスタリングを行い，デンドログラムを順次作成していく手法も考えられるが，今回はインタラクティブ性において重要である描画速度を重視するため，事前にクラスタリングを行っておく手法をとる．本手法のグラフはノード，エッジ，クラスタの集合の構造になっている．描画すべき要素を判断するため，3 つの集合の要素ははそれぞれ,「Active」「nonActive」の値を持つようにした．t プログラム

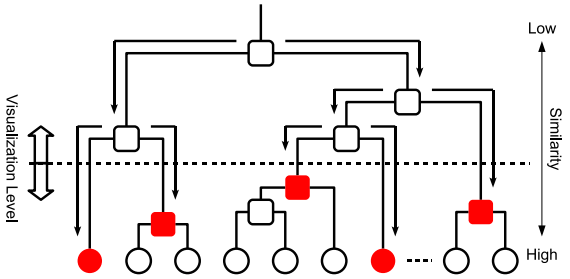


図 4. スライダ変更時の走査

の起動時とスライダの値 t が変更された場合，以下の順で各要素の「Active」「nonActive」を切り替えていく．なお，初期値は nonActive とする．

1. 探索クラスタ c = ルートクラスタとおく
2. $S(c) < t$ であれば
   (a) c の保持しているノードすべてを Active とする
   (b) c を nonActive とする
   (c) c が子クラスタを持っていたら，c = 子クラスタとし 2 に戻る（子クラスタが複数あった場合はすべての子クラスタについて探索を行う）
3. $S(c) \geq t$ であれば
   (a) c の保持しているノードすべてを nonActive とする
   (b) c を Active とする

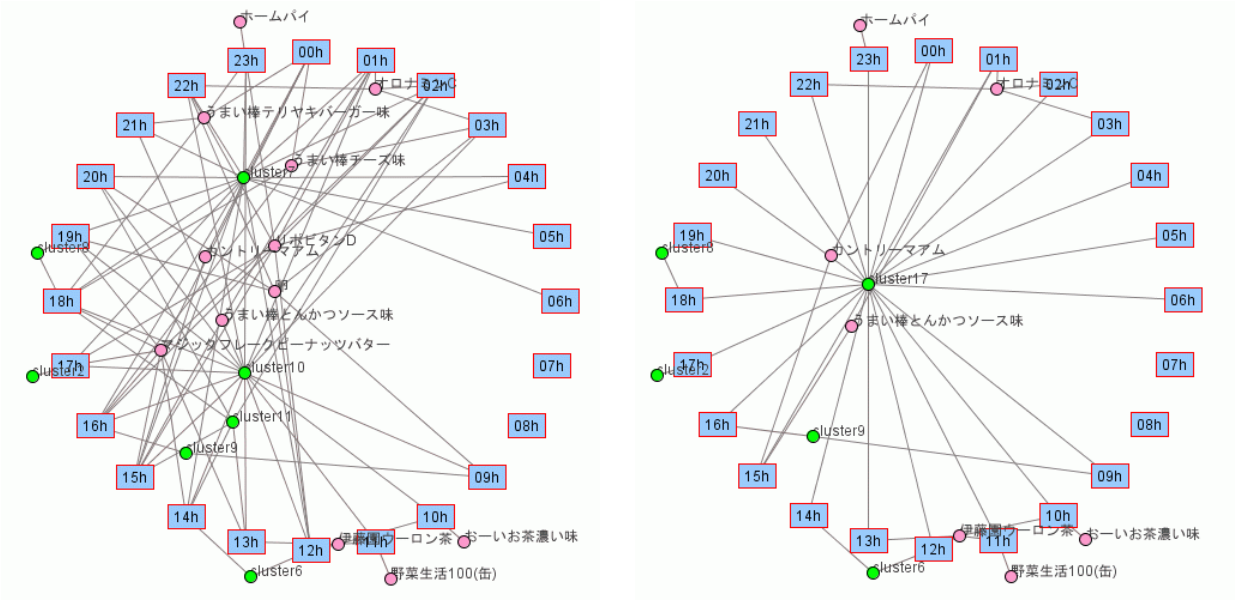
走査が終了したら，Active となっているノードとクラスタを接続するエッジのみを Active とする．最終的に，Active となっている，ノードとエッジ，クラスタでグラフを構成することで，要求を満たすグラフを可視化することが可能となる．

図 4 にクラスタリングの構造を樹状図で示す．クラスタリングの構造は木構造であらわされ，ノードを円，クラスタを四角として表示している．スライダが動かされると図中のグラフを水平に横断している点線部が上下に移動する．クラスタの探索の動きを矢印で，それにより決定された Active なクラスタとノードのみ塗りつぶして表示している．

3.3 等類似度線描画

グラフを縮約させる手法とは別の観点で可読性を向上させる手法として，読み手の視点を注視させる手法の導入が考えられる．そのアプローチとして，等類似度線描画を提案する．

人間の注視特性より，重要な部分を囲み線などであらわす方法が有効であることは良く知られている．クラスタリングを用いない可視化結果からもクラス



(a) クラスタリング（閾値=0.50）　(b) クラスタリング（閾値=0.35）

図 3. クラスタリングの効果

タの形成を読み取ることは出来るが，クラスタリングの結果を等高線のように表示することでノード間の関連性を能動的に注視させ，より理解しやすくさせるのが狙いである．

インタフェースは，クラスタリングスライダのインタフェースと同様スライダの形式を取っており，インタラクティブに表示する囲み線の類似度の閾値を変更することが可能である．スライダの値も同様最大値 101[%]，最小値 0 を設定している．スライダの値が 101% に位置しているときには囲み線は表示されず，0 に行くに従い増えて行き，最終的に全ノードが類似度線によって表現される．

実装方法

本研究で用いるクラスタリングは木構造を持っているため，領域の入れ子構造で表現することが可能である．探索方法はクラスタリングスライダとほぼ同様であるが，Acitve，nonActive の切り替えはクラスタのみに適用する．

1. 探索クラスタ c = ルートクラスタとおく．
2. $S(c) < t$ であれば，c を nonActive とする
3. $S(c) \geq t$ であれば，c を Active とする
4. c が子クラスタを持っていたら，c = 子クラスタとし 2 に戻る．(子クラスタが複数あった場合はすべての子クラスタについて探索を行う)

走査が終了したら，Active とされているクラスタを，そのクラスタが保持しているノード（子クラスタが保持している分も含め）を囲む閉曲線を描画する．閉曲線描画は以下のステップで行う．

1. クラスタに含まれるノード集合を V_A とする．
2. ノード集合 V_A の最外周凸多角形を形成するノード集合 V_B を探索する（Graham 走査 [6]）
3. ノード集合 V_B の周囲を直線と円弧による閉曲線で囲む．

図 5 は閾値 $t = 0.53$ で類似度線を表示した例である．等高線のように，等しい類似度のノードが囲まれ，関連度の強いノードを瞬時に理解することが出来る．

4 議論

クラスタリングスライダの狙いは，クラスタリングの強さをインタラクティブに変更することで，読み手に適切なノード量で表示することであった．利用して得られた知見として，ノードの関連性を読みとることが出来ることである．一例として，初期ノード配置では離れている「うまい棒」の味違いは，スライダを動かしていくと他のノードより比較的早い段階で 1 つに結合するが，図 3(b) から分かるとおり，「とんかつソース味」は閾値 0.35 の状態でもノードとして残っている．すなわち，ここから「うまい棒」の購買傾向は似ているが，「とんかつソース味」だけは購買傾向が異なり，他の味と比較して 15 時に買われる傾向が強いことが認識できる．

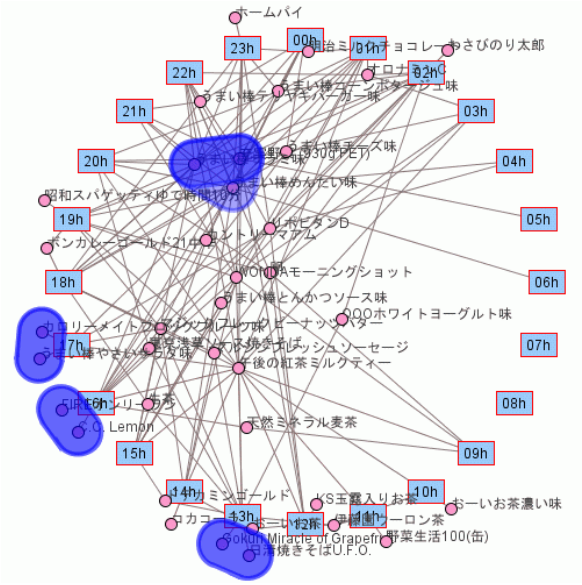


図 5. 等類似度線描画（閾値=0.53）

スライダを左の向きに操作し結合するノードの類似度の閾値を下げていき，ある一定の値でノードが結合した際，スライダを右に一旦戻す動作をすると今集まったノードがどれかを確認することが可能となる．ただし，人間が能動的に確認作業をする必要があり，システム側がそれを可視化した方が良いとも考えられる．

加えて，同じ類似度で結合するクラスタが複数あった場合は，それぞれのクラスタの要素を確認することが難しい．例として，10 個のノードが 2 つのクラスタに結合した場合，片方のクラスタは 10 個中どの要素が集約されたかは集約される前の位置により判断するしかない．そのため，クラスタの集約の過程の表現については検討の必要がある．

等類似度線の描画についても，クラスタリングスライダを動かしノードを集約し解放する性質と同様のものを持っていると感じられた．違いとしては，ノードが結合せず残っているため，ある一瞬の静止画でノードの関連性が読みとることが可能で，一覧性の性質が強いことである．

今回の手法では，類似度の囲み線を表示するのみにとどまっているため，ノードの配置に依っては類似度線が重なってしまう．そのため，関連度の強いノードは近くに配置する等のレイアウトと絡めた工夫が必要である．

一部のノードとクラスタの関係性を知るにはノード集約を行うクラスタリングスライダを，全体のクラスタの分布から関連性を知るには等類似度線描画を切り替えて使うことで適切な情報を可視化し，人間の認識と理解を促進することが可能である．

5 関連研究

クラスタリングを事前処理として行い可視化結果に反映させる手法として，小林らの研究がある [7]. Web における情報検索の可視化を行うため，クラスタリングを事前処理として行い可視化結果に反映させる手法を用いている．

クラスタリングの情報を描画する手法として，井上らの研究がある [8]．クラスタリングの結果を可視化するため，樹状図を描画レベルと注視するノードを変更可能にしたネットワーク図で描画する手法を用いている．

クラスタ構造のネットワークを描画する手法として，Frishman らの研究がある [9]．クラスタ構造を持つグラフを二次元もしくは三次元平面上に描画し，グラフの情報の動的変化に対応した自動レイアウトを行うことが特徴である．

6 まとめ

大規模 2 部グラフの可視化における可読性向上について，いくつかのインタラクティブな可視化手法の有効性について述べた．従来の描画手法に加え，階層型クラスタリングとその操作を用い，全体の俯瞰から詳細の情報まで一括のインタフェースで取得できる，可読性に優れたグラフ描画ツールを構築した．

参考文献

[1] K. Misue. “Drawing Bipartite Graphs as Anchored Maps,” In *Proc. of APVIS2006*, pp.169–177, 2006.

[2] P. Eades. “A heuristic for graph drawing,” Congressus Numeranitium, Vol.42, pp.149–160, 1984.

[3] 杉山 公造． グラフ自動描画法とその応用. 計測自動制御学会, 1993

[4] J. A. Hartigan. クラスター分析, マイクロソフトウェア, 1983.

[5] 神嶌 敏弘． “データマイニング分野のクラスタリング手法 (1) — クラスタリングを使ってみよう！—,” 人工知能学会誌, vol.18, no.1, pp.59–65, 2003.

[6] R.L.Graham, “An efficient algorithm for determining the convex hullof a finite planar set,” Information Processing Letters, vol.1, 132–133, 1972.

[7] K. Kobayashi et al. “Information Gathering Support Interface by the Overview Presentation of Web Search Results,” In *Proc. of APVIS2006*, pp. 103–108, 2006.

[8] 井上 悦子ら. “大規模クラスタリング結果のグラフによるインタラクティブな可視化手法,” 情報処理学会 情報学基礎研究会報告, 2006-FI-85-(4), pp. 21–28, 2006.

[9] Y. Frishman, A. Tal. “Dynamic Drawing of Clustered Graphs,” In *Proc. of InfoVis2004*, pp.191–198, 2004.