① 配置手法

隣接性と特徴量に基づくクラスタリング手法を導入したグラフ可視化

伊藤　貴之○（お茶大）

Graph Visualization Applying a Cluste with Combination of Adjacency ar

Takayuki ITOH

昨日の面談の際にお伝えした、村上さんの研究の元論文を紹介します。

まずは階層型グラフの配置手法についてです。
村上さんは以下の論文
Key-node-Separated Graph Clustering and Layout for Human Relationship Graph Visualization
http://www.is.ocha.ac.jp/~itot/paper/ItotRJPE25.pdf
に載っている手法を使っています。
この手法は伊藤が自分で実装したものでJavaで書かれています。
日本語4ページに短縮した論文を添付します。
（隣接性と特徴量に…というやつです）

続いて階層型グラフの評価手法ですが、以下の論文になります。
The sprawlter graph readability metric: Combining sprawl and area-aware clutter
こちらはPDFにアクセスできないので直接添付します。
こちらも伊藤がJavaで実装しています。
日本語4ページに短縮した論文も添付します。
（空間浪費度と画面乱雑度に…というやつです）

村上さんの実装は上述の2種類の手法をPythonのコードで反復的に呼び出しています。
このうち配置手法のほうを複数組み合わせてはどうか、という話になります。

まずは勉強してみてください。

ABSTRACT

Graph visualization has evolved with a variety of applications information can be represented as graphs. Many graph drawing method based on density of edges to find tightly connected subgraphs, and clustered graphs. On the other hand, it is not always preferable if users (called key nodes in this paper) but they are hidden in the large cl effective to separately visualize the key nodes detected based on conn This paper presents a graph drawing technique for attribute-embedded g a graph clustering algorithm taking into account the combination of con clustering step divides the nodes according to the commonality of c feature value vectors. It then calculates the distances between arbitrary number of connected edges and similarity of feature value vectors, an on the distances. Consequently, the technique separates important no multiple large clusters, and improves the visibility of connections o examples with human relationship graph datasets, including a co-auth network datasets.

Keywords: Graph visualization, graph clustering.

1．序　論

近年のグラフ可視化手法の多くは，グラフを構成するノードに前処理としてクラスタリングを適用してからクラスタ単位でノードの配置を算出する．そしてクラスタリング手法の多くとして，エッジの密度の濃い部分をクラスタ化するアルゴリズムが導入されている．このような手法はコミュニティ発見などの多くの目的で有用である反面，多くのノードと連結された重要なノードがクラスタの中に埋もれて目立ちにくくなるといった問題がある．Fig.1 にその端的な一例を述べる．Fig.1(a)において赤いノードは他の多くのノードと連結されている．しかし一般的なクラスタリング手法を適用すると，Fig.1(a)(1)に示す大きなクラスタに赤いノードが包括される可能性が高い．このようなクラスタリング結果を Fig.1(b)のように描画すると，赤いノードはクラスタに埋もれて発見しにくくなる．多くのアプリケーションにおいて，このように多くのノードに連結された重要なノードは，Fig.1(c)に示すようにクラスタから独立されるような形で扱われることが望ましい．そうすればクラスタを単位として描画した Fig.1(d)をみても，赤いノードが他の多くのクラスタを連結させる重要な役割を果たしていることが容易に視認可能となる．

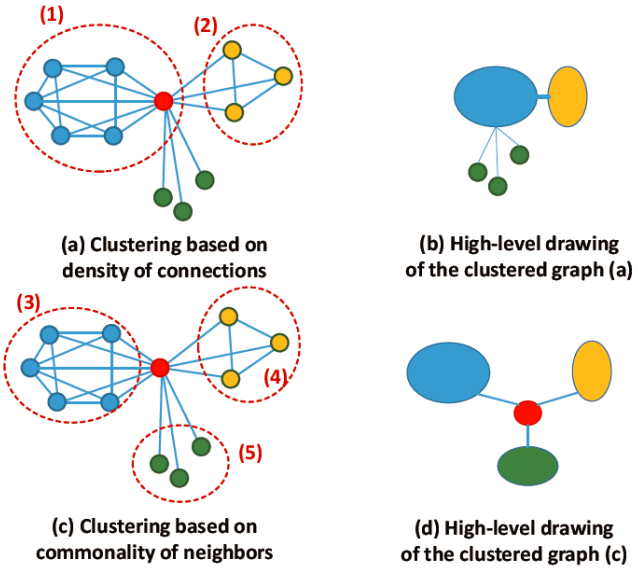

Fig.1　グラフクラスタリングを適用した可視化

このようなグラフ可視化を実現するために本報告では，

隣接性と特徴量に基づくクラスタリング手法を導入したグラフ可視化の試みについて報告する．本手法では以下の 2 つの基準によってノード間距離を定義する．1 つめはエッジによって連結された隣接ノードの共通性，2 つめはノードに付与された特徴量ベクタの類似性である．この距離にもとづいてノードをクラスタリングすることにより，重要なノードが独立された形でのグラフ可視化を実現できる．

Fig.2 に提案手法による可視化例を示す．Fig.2(上)に示すように本手法は，1 つのクラスタを構成する複数のノードが同心円上に放射状に配置する．ここで，表示されるべきエッジの本数，また同一クラスタ間に生成されるエッジを束化する条件，などを対話的に修正できるようになっている．また Fig.2(下)に示すように本手法では，ユーザが特定のノードをカーソルで指すことで，そのノードに直接連結されたエッジを選択表示できる．

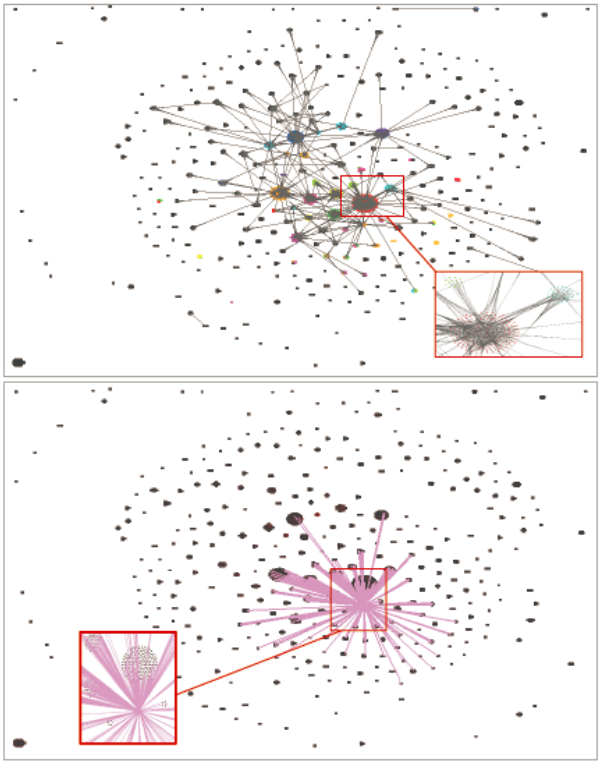

Fig. 2　提案手法による可視化例

2．関連研究

グラフクラスタリングに関する手法は既に多数報告されており，Schaeffer[1] はその網羅的なサーベイ論文を発表している．最も多くの場面に用いられているクラスタリング手法として，多数のエッジによって密接に接続された部分グラフを抽出する手法が知られている．Schaeffer のサーベイには他にも多くのクラスタリング基準が紹介されており，本報告が採用する「エッジによって連結された隣接ノードの共通性」と「ノードに付与された特徴量ベクタの類似性」もそれに含まれる．

重要なノードを強調的に表示するという発想はグラフ可視化において重要である．大澤らが発表した KeyGraph[2] は部分グラフを連結するノードを事前発見して強調表示する．Correa ら [3]の手法ではセントラリティという基準値を利用することで，ソーシャルネットワークにおける重要人物を強調表示している．Hong ら [4] の手法では二部グラフを採用することで，重要なノードに目が届きやすくなるようなグラフ配置を実現している．これらの手法と違って本手法では，例えば力学モデルや次元削減などを採用した汎用性の高いグラフ配置手法を踏襲する形で，重要なノードを強調的に表示するようなグラフ可視化手法を実現する．

本報告が提案するようなクラスタ単位のグラフ配置手法は既に多く報告されている．本報告ではバネモデル [5] またはストレス最小化モデル [6]によるクラスタ初期配置に対して Delaunay 三角メッシュのスムージング処理を施すという形でグラフを配置しているが，それ以外の手法を適用することも可能である．例として，階層型データ可視化のための空間充填アルゴリズムを搭載したグラフ配置手法手法 [7][8]を採用してもよい．

3．提案手法

本章では本報告の提案手法について，データ構造および処理手順を説明する．

3.1　データ構造

提案手法は入力データとして無向グラフを想定する．また各ノードには多次元ベクタとなる特徴量が付与されているものとする．また入力データ中の 1 個以上のノードの集合をクラスタと呼ぶ．本手法ではクラスタを単位として各ノードの画面上の位置を算出するものとする．

3.2　グラフクラスタリング

提案手法では任意の 2 ノード間距離を算出し，その距離にもとづいてノードをクラスタリングする．現時点の著者の実装では，クラスタリング手法として最長距離法に基づいた階層型クラスタリングを適用している．また提案手法ではノート間距離として以下の 2 種類の距離

d_{vec} ： 特徴量ベクタに基づくノード間距離

d_{adj} ： 隣接ノードの共通性に基づくノード間距離

を算出し，これの一次結合を表す以下の式

$$d = \propto d_{vec} + (1-\propto) d_{adj}$$

によりノード間距離 d を算出する．以下，2 種類の距離の定義を述べる．

特徴量ベクタに基づくノード間距離

2 個のノードに付与された特徴量ベクタをa_iおよびa_jとする．このとき本手法では

$$inner = a_i \cdot a_j / |a_i||a_j|$$

により特徴量ベクタの内積を求める．この値を用いて以下の式

$d_{vec} = 1.0 - \text{inner}$

によってノード間距離を算出する．

隣接ノードの共通性に基づくノード間距離

2 個のノードの両方にエッジで接続されている隣接ノードの個数をn_{adj}とする．この値を用いて以下の式

$d_{adj} = 1.0/(1.0 + n_{adj})$

によってノード間距離を算出する．

3.3 ノード配置

提案手法では階層型クラスタリングによるクラスタ生成ののち，クラスタを単位として各ノードの配置を決定する．現時点での我々の実装では，以下のアルゴリズムによってノードの画面上の位置を算出する．

1. クラスタをノードに置き換えたグラフを生成し，これに対して一般的なグラフ配置アルゴリズムを適用する．現時点で我々は以下の 2 種類のグラフ配置を試している．
   a) エッジにバネの力学モデルを適用した Spring-force モデルによる配置計算方法．
   b) PivodMDS という次元削減手法による初期配置およびストレス最小化モデルによる配置計算方法．
2. 各クラスタを構成するノード数から，各クラスタの半径を計算する．
3. クラスタをノードに置き換えたグラフの配置結果から，各ノードを接続する Delaunay 三角メッシュを生成し，このエッジ長が両端のノードの半径の合計値にできるだけ近づくように三角メッシュに対してスムージング処理を適用する．
4. 2.で求めた半径にしたがって各クラスタを円に置き換え，その円内部にクラスタを構成する各ノードを放射状に配置する．

3.4 エッジ束化

我々の実装では同一クラスタペアに属する 2 ノードを連結するエッジ群を束化することにより表示を単純化している．クラスタリング等によってノードが階層化されたグラフにおいてエッジを束化する手法は数多く報告されているが，本手法では階層が 1 段だけであることを前提に，以下の単純化された実装を採用している．

Fig.3 に我々の実装におけるエッジ束化の処理手順を示す．ここでは Fig.3(a)のように，青いノードが示すクラスタと，緑のノードが示すクラスタを連結するエッジ群を束化することを考える．我々の実装では Fig.3(b)(左)のように，クラスタの中心点を連結する線分上に 2 点を配置し，これを制御点とする Bezier 曲線を生成することで，擬似的に束化されたようにエッジ群を描画する．あるいは Fig.3(b)(右)に示すように，2 ノードを連結する線分と，クラスタの中心点を連結する線分について，それぞれ 3 分割した点を生成し，これらの点を連結する線分上に 2 点の制御点を生成する．Fig.3(b)(左)および Fig.3(b)(右)における灰色の円が，制御点に相当する．

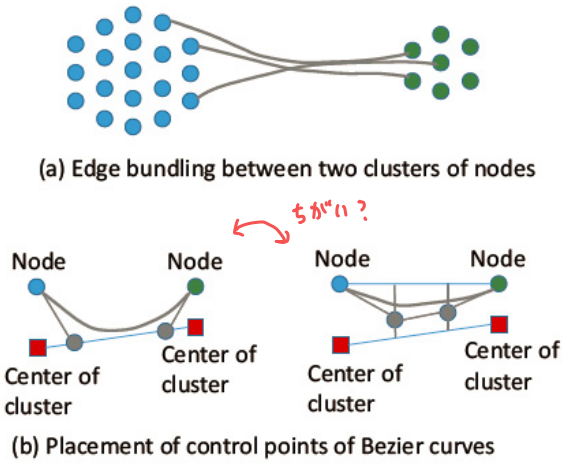

Fig. 3 エッジ束化の単純な実装

3.5 ノード描画

ノード描画において色や高さを与えることが一定の効果をもたらす．我々の実装では以下を搭載している．

[色付け方法 A] ノードに付与された特徴量ベクタの各次元に固有の色を与え，各ノードに対して最も値の大きい次元の色を与える．

[色付け方法 B] ノードの重要度（例えば連結されたエッジの本数）に応じて強調的に色や高さを与える．

また巨大なグラフにおいては，個々のノードを描かずにクラスタを 1 個の円として描くことも効果的である．

4．適用事例

本章では適用事例として，NBAF (NERC Biomolecular Analysis Facility) が 1998-2013 年に発表した 564 本の論文群に対して，著者をノード，共著関係をエッジとしたグラフを可視化した事例を示す．このグラフにおいてノード（著者）は 1,821 個，エッジ（共著関係）は 11,097 本であった．

まず我々は著者の特徴量を計算するために，各著者が発表した論文タイトルから単語の出現頻度を集計し，頻度の高い50単語を抽出した．その中から「実験」「アルゴリズム」といった汎用性の高すぎる単語を削除するなどして，当該専門分野を象徴する20単語を主観的に選択した．その上で各単語の著者ごとの出現頻度を再集計し，より多くの著者のタイトルに出現する12単語を採用した．以上の処理により著者が選んだ12単語と色の関係は以下の通りである．Genetic (赤), Molecular (橙), Loci (黄), Microsatellites (黄緑), Isolation (緑), Inbreeding (青緑), Transcriptomics (水色), Expression (青), Bacterial (藍色), Breeding (紫), Polymorphic (ピンク).

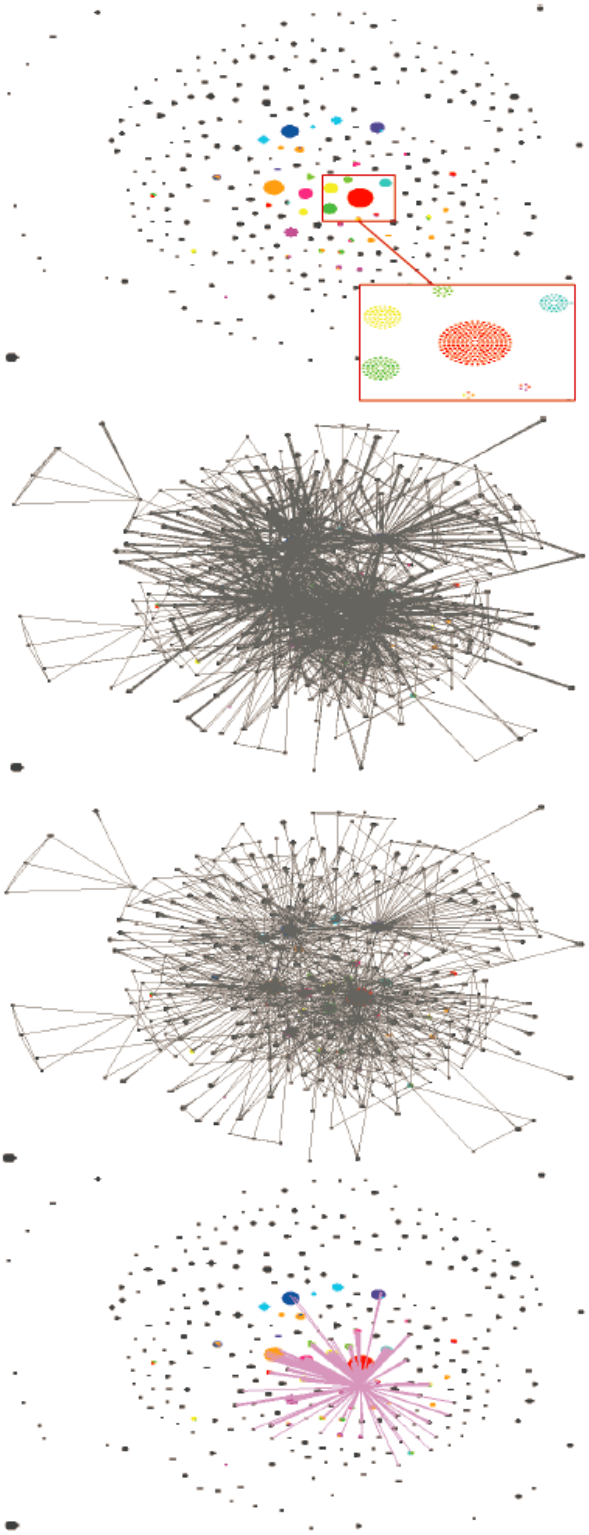

Fig. 4　可視化例．上から(1)レイアウト結果，(2)エッジ束化適用前，(3)エッジ束化適用後，(4)ユーザ指定ノードに連結されたエッジのみの表示．

図4に可視化結果の例を示す．数個～数十個のノードで構成されるクラスタが円形に表示され，画面内で適切な距離を保って配置されているのがわかる．またエッジ束化によってエッジ分布の視認性が向上しているのがわかる．同一データを10回連続で可視化した際の平均計算時間は，2.6GHzデュアルコアCPUを搭載したWindows 7 (64ビット) のPCにおいて，グラフクラスタリングに2.5秒，クラスタ配置に7.3秒，メッシュスムージング処理に1.5秒であった．

図5は本手法が搭載しているクラスタリング手法と一般的なクラスタリング手法（本報告ではモジュラリティにもとづいてエッジの密な部位を抽出する手法）を比較した結果である．図5において各ノードの色は連結されたエッジの数に比例して赤に近づいている．図5(上)では最も赤い2個のノードが大きなクラスタから分離して小さなクラスタに属しているのに対して，図5(下)では赤い2個のノードが大きなクラスタに属していることがわかる．よって本手法が搭載するクラスタリング手法のほうが，本研究の目的に近いクラスタリング結果をもたらしていることがわかる．

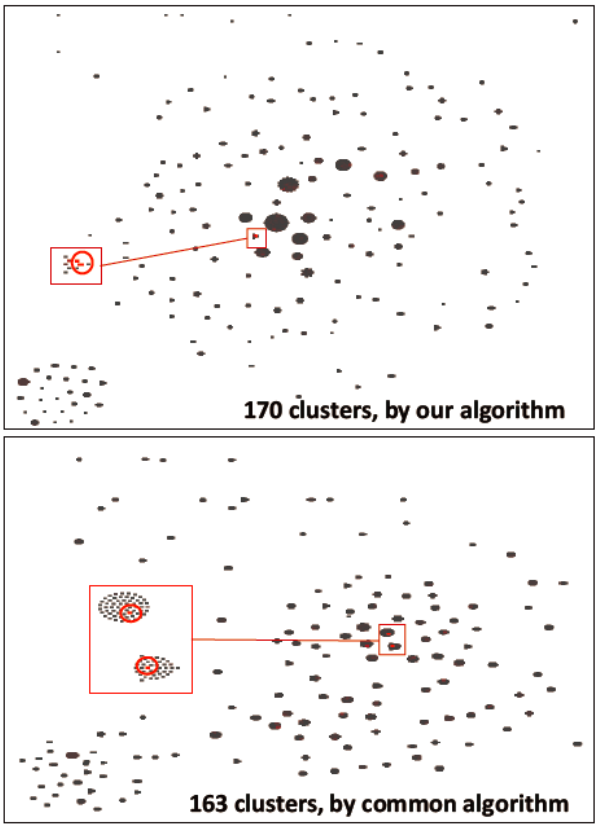

Fig. 5　(上)本手法が採用するクラスタリング手法による結果．(下)一般的なクラスタリング手法による結果．

続いてこの可視化結果から論文著者群についてどのような傾向が観察できるかを解説する．今回用いたデータをいろいろな手段で可視化した結果を図 6 に示す．図6(1)では各ノードを[色付け方法 A]で描画している．この結果からわかるように，大きなクラスタの多くは単一色で構成されている．言い換えれば，それぞれのクラスタは強力な専門用語によって構成されていることがわかる．

続いてノードの色を[色付け方法 B]に切り替えた結果を図 6(2)に示す．この結果において，赤く表示されたノードが重要度の高いノードである．赤いノードの各々をクリックして，それらから連結されたエッジを観察した結果を以下に示す．

図6(3)(4)は，図6(2)で赤く表示されたノードのうち同一クラスタに属する2個のノード，著者名でいうと T. Burke および D. A. Dawsonの各々に接続されているエッジを表示した例である．この2名は同一クラスタに属するだけあって，非常に類似した共著関係を有することがわかる．一方で，この可視化結果を細かく比べると小さな差異がいくつか見受けられる．T. Burkeのほうが少数のクラスタと強い接続関係を有しており，Geneticなどの特定の用語でその共著関係が成立していることがわかる．一方でD. A. DawsonはT. Burkeよりもやや広い共著関係を有しており，その中にはLoci, Inbreeding, IsolationといったT. Burkeの論文には見られない用語が関係していることがわかる．

図6(5)(6)は，図6(2)で赤く表示されたノードのうち別の2個のノード，著者名でいうとA. R. Cossins および J. K. Chipman の各々に接続されているエッジを表示した例である．A. R. Cossins は用語でいうとMolecular, Expression, Bacterial などに関係した共著関係を多く有しており，前述のT. Burke および D. A. Dawson とは異なる専門分野でのキーパーソンであることが見受けられる．一方で J. K. Chipman は用語でいうと Transcriptmics, Expression などに関係した共著関係を多く有していることから，前述の3人のいずれとも異なる専門分野でのキーパーソンであることが見受けられる．また他の3人と比べて灰色のノードと接続されたエッジが多いことから，この著者の専門分野を可視化結果から読み取るには現時点で選出されている12単語以外の用語も追加選出する必要があることが示唆される．

５．まとめ

本報告では，隣接ノードの共通性と特徴量ベクタの類似性に基づいたクラスタリング手法を適用したグラフ可視化の一手法を提案した．本報告が適用したクラスタリングにより，多くのクラスタと接続関係を有する重要なノードが大きなクラスタから分離されるような構造を得られやすくなり，結果として重要なノードとその周辺のノードとの接続関係を視認しやすくする．また本報告では論文著者をノード，共著関係をエッジとしたグラフを例題として，その可視化結果から本手法の効果を検証した．

今後の課題として，提案手法によるグラフ可視化結果の定量的および主観的な評価を進めるとともに，論文共著関係グラフ以外のデータについても適用してその校歌を検証したい．また本報告では無向グラフを対象として処理手順や適用事例を示してきたが，有向グラフの可視化に特有の問題についても議論を進めたい．

参 考 文 献

1) S. E. Schaeffer, Graph Clustering, Computer Science Review, 1(1), 27-64, 2007.
2) Y. Ohsawa, N. E. Benson, M. Yachiba, Key-Graph: Automatic Indexing by Co-occurrence Graph based on Building Construction Metaphor, IEEE International Forum on Research and Technology Advances in Digital Libraries, 1998.
3) C. Correa, T. Crnovrsanin, K.-L. Ma, Visual Reasoning about Social Networks Using Centrality Sensitivity, IEEE Transactions on Visualization and Computer Graphics, 18(1), 106-120, 2012.
4) S. Hong, W. Huang, K. Misue, W. Quan, A Framework for Visual Analytics of Massive Complex Networks, International Conference on Big Data and Smart Computing (BIGCOMP), 15-17, 2014.
5) P. Eades, A Heuristics for Graph Drawing, Congressus numerantium, 42, 146-160, 1984.
6) E. R. Gansner, Y, Hu, S. North, A Maxent-Stress Model for Graph Layout, IEEE Transactions on Visualization and Computer Graphics, 19(6), 927-940, 2013.
7) T. Itoh, C. Muelder, K.-L. Ma, J. Sese, A Hybrid Space-Filling and Force-Directed Layout Method for Visualizing Multiple-Category Graphs, IEEE Pacific Visualization Symposium, 121-128, 2009.
8) W. Didimo, F. Montecchiani, Fast layout computation of clustered networks: Algorithmic advances and experimental analysis, Information Sciences, 280, 185-199, 2014.

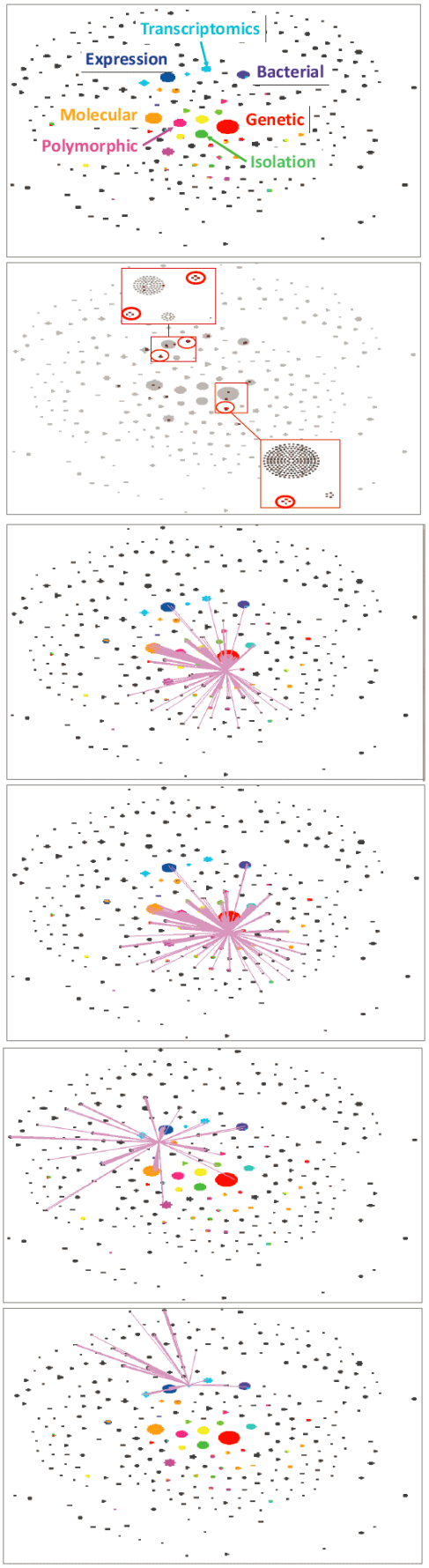

Fig. 6　論文共著関係グラフにおける適用事例.