

カテゴリ付きネットワークの可視化手法と適用事例

理学専攻 情報科学コース 中澤 里奈 (指導教員：伊藤貴之)

1 概要

近年、カテゴリ付きネットワークに関する様々な可視化手法が提案されている。本論文では、カテゴリ付きネットワークにおける、カテゴリ毎に束ねることで大量のエッジ情報を要約する可視化手法を提案し、遺伝子ネットワークを対象とした無向ネットワークの適用事例と、論文参照関係を対象とした有向ネットワークの可視化事例を示す。本要旨では紙面の都合上、適用事例のみを紹介する。

2 遺伝子ネットワークへの適用事例

遺伝子の機能や他の遺伝子との関係性の解明のための情報共有手段として、各遺伝子をノード、遺伝子間相互作用をエッジとする遺伝子ネットワークが利用されている。遺伝子ネットワークの可視化はこれまでも試みられているが、既存手法には以下のような問題点が残っている。

- (1) ノード数やエッジ数が非常に多いため、その中でどこが注目すべきか一目で理解するのが難しい。これを解決するためにはネットワーク構成を明快に表示することが重要である。
- (2) エッジの接続構造が複雑であるため、画面上での重なりが多く生じる。これが機能と相互作用の視認性を妨げることがある。

そこで遺伝子ネットワークと遺伝子機能情報に対し本手法を適用する。

2.1 提案手法

遺伝子機能情報の分類 本手法では遺伝子機能情報として、Gene Ontology(GO)[2]を用いる。GOは遺伝子の機能や概念を共通語で記述するための用語体系およびその策定プロジェクトで、機能が解明された遺伝子にはGOが規定する1個以上の用語(GOターム)が割り当てられている。本手法では、より多くのGOタームから得られる知見を可視化結果に反映するために、もともとDAG構造をとっているGOタームを、データ構造上の距離をもとにクラスタリングする。まずGOから1個のGOタームを選び、それと直接接続される子のGOタームを全て抽出する。これらのGOタームが割り当てられている遺伝子数の合計が一定値以上であるものを、1個のクラスタとする。このようにして生成されたクラスタ群を、最上位の親ノードによってグループ化し、1個のグループに属するクラスタ群のうち10~15種類のGOタームクラスタをアイテム情報としてノードに付与し、可視化に用いている。

遺伝子クラスタリングと配置 各遺伝子に割り当てられたGOターム群を参照して、それらが属するGOタームクラスタを各遺伝子に割り当てる。続いて、各遺伝子をGOタームクラスタの組み合わせで分類し、全く同一のGOのタームクラスタ群を有する遺伝子群をまとめて1個のクラスタとする。そして、クラスタを単位としてFRUITS Net[1]の配置アルゴリズムを用いることで、遺伝子に対応する各ノードの配置を決定

する。FRUITS Netの配置アルゴリズムは力学モデルと空間充填モデルの2ステップで構成される。まず遺伝子クラスタをノードとするグラフ構造を形成し、これに力学モデルの反復計算を適用することで、遺伝子クラスタの仮の位置を決定する。これによって、エッジで連結された遺伝子クラスタや、同一のGOタームクラスタを共有する遺伝子クラスタを、画面上で近くに配置する。続いて、この仮の配置結果を参照しながら空間充填モデルを適用することで、遺伝子クラスタの画面上での重なりを回避しつつ、配置画面占有面積を減らす。

エッジバンドリング処理 続いて、エッジの視認性を向上するために束化を施す。まず一つの遺伝子クラスタに注目し、他の遺伝子クラスタとの間に接続しているエッジの本数を集計する。遺伝子クラスタ間のエッジが所定の本数以上ならば、クラスタ間に太線を束として描き、接続していたエッジとその束を結び、さもない場合は、エッジをそのまま描く。束がノードを横切ってしまうと視認性がさがるので、遺伝子クラスタを回り込むように束の折れ線を生成する。

2.2 可視化結果

本論分はiRefIndex[3]に公開されているショウジョウバエの遺伝子ネットワークを適用し、各遺伝子をノード、タンパク質間相互作用をエッジとし、GOによって定義される遺伝子機能情報をノードに付与して可視化を試みた。ノード数は8945、エッジ数は32703、既知のGOターム数は259、今回アイテム情報としてノードに付与したGOタームクラスタは12個であった。実行結果を図1に示す。図1の中央部から右部にかけての機能間の主要な関係性が視認しやすくなり、ここでは(a)と(b)、(a)と(c)、(b)と(e)、(d)と(e)の各遺伝子クラスタ間において、より多くの相互作用が見られることがわかった。本手法を適用することで視認性が向上し、より効果的に遺伝子ネットワークの関係性を把握し、機能の推測ができるようになったと言える。

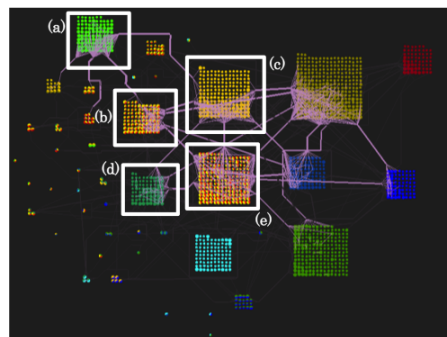


図1: 遺伝子ネットワークの可視化結果例

3 論文参照関係への適用事例

研究を進めていく際の文献調査方法は、検索エンジンを用いたり、既に読んでいる論文の参考文献をたどっていったりするなどがある。しかし、研究を始めたばか

りの学生は検索のための適切なキーワードを全て知っているわけではないため、検索エンジンによる方法では、検索ワードによって見落としがおこってしまう上に、論文同士の関係を見いだすことができない。そこで本手法では、研究分野と参照関係からの文献探しに適した参照関係可視化手法を提案する。

3.1 提案手法

文献のカテゴリ分類 本手法では、文献の内容から文献を研究の分野別にカテゴリ分類するために、今回は一例として潜在的ディリクレ配分法 (LDA) [4] を用いた。LDA は、一つの文書に対して複数のトピックの存在を想定した、文書生成過程の確率的なトピックモデルである。LDA で推定されるトピックを一つの研究分野と見なし、各文献がどのような研究分野に属するかを分類する。LDA において、トピック数は予め決めておく値であるため、今回はトピック数を 12 とした。

ネットワーク配置とエッジバンドリング処理 続いて、LDA を用いて推定したカテゴリ分類をもとに、FRUITS Net[1] の配置アルゴリズムを用いることでノードの配置を決定する。これにより、参照関係がある文献同士を近く、同じ研究分野に属する文献同士を近く配置し、ノード同士の重なりを回避しつつ、画面配置占有面積を低減することができる。ノードの大きさはその文献の被引用数に比例する。その上で、エッジの視認性を向上するためにエッジバンドリングを施すことで、大量の線情報を束ね、要約する。ユーザのスライダ操作によってエッジを束ねる本数の閾値を設定できる。カテゴリ分類によってできた文献クラスタ間のエッジ本数を計算し、それが閾値以上であるならば、本数に比例した太さの束として描く。閾値に満たない場合はエッジをそのまま描画する。

ユーザインタフェースでの描画と操作 ユーザは操作パネル上で可視化結果の拡大縮小や平行移動操作、エッジバンドリング適用の有無、その閾値の設定を行うことができる。また、各文献であるノードがクリックされると、その文献のタイトルや発行年、概要などの情報を表示する。同時に、クリックされたノードが持つエッジと、それに接続するノードが持つエッジの二世代分の参照関係をハイライト表示する。このエッジのハイライトは 2 つのノードについて同時に適用可能であり、これによって両ノードの接続関係を比較することができる。一方、最初に注目すべき文献を探すことは初学者にとっては容易ではない。そこでユーザは大まかな研究分野かキーワードを利用することで、画面上に表示させる文献をフィルタリングすることができる。自分の興味のある研究分野カテゴリを操作パネル上のボタンから選択した場合、選択した研究分野のみと分類された文献群をウィンドウ中央部に拡大表示される。また、キーワードを入力すると、入力された語がタイトル中に含まれている文献ノードのみが表示される。

3.2 可視化結果

使用した文献データは 1990 から 94 年、2000 から 2010 年までに SIGGRAPH で発表された論文 1072 本である。例として、ユーザが写真や動画の処理関連の

研究の文献を調査したい場合を考える。写真、動画関連の研究分野カテゴリを選択すると、表示結果は図 2 のようになる。図 2 の (A) が写真、動画の処理関連の分野であり、(A) に接続する束を追っていくと、(A) には (B) と (C) のどちらかと参照関係を持つ文献が多く存在するということがわかる。(B) は大域照明、光反射や光拡散モデルなどと写真や動画の処理が組合わさったクラスタ、(C) は画像処理関連のクラスタである。この結果から、写真や動画の処理に関する文献を探したい場合、優先して探したい内容がどちらかという (B) の照明関連寄りなのか (C) の画像処理技術に関連寄りなのかによって、その分野と接続している文献から先に読み進めていけばいいということがわかる。従って、本インタフェースを利用することで関係ありそうな論文を研究分野と参照関係から整理して、どの論文群から先に読むかを選択できるのではないかと考えられる。

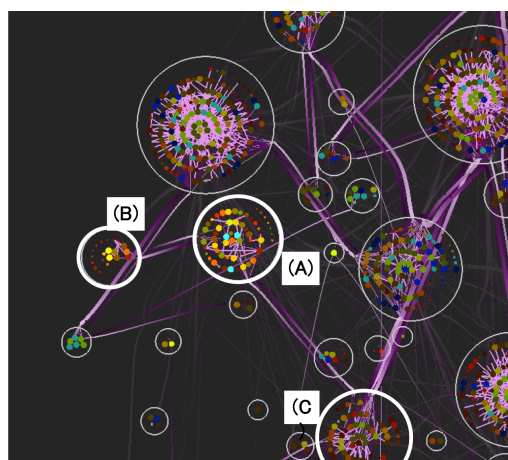


図 2: 論文参照関係可視化結果

4 まとめ

本論文では、カテゴリ付きネットワークデータの可視化手法を提案し、2 つの適用事例を紹介した。

謝辞

遺伝子データへの適用に際し、データ及び貴重なご意見をいただきました東京工業大学の瀬々 潤准教授、寺田 愛花氏、結果の考察に際し貴重なご助言をいただきました産業総合研究所の油谷 幸代氏、論文参照データをご提供いただきました東京農工大学の斎藤 隆文教授。LDA に関してご助言をいただきましたお茶の水女子大学の江里口 瑛子氏に感謝いたします。

参考文献

- [1] T. Itoh, C. Muelder, K. Ma, J. Sese: A Hybrid Space-Filling and Force-Directed Layout Method for Visualizing Multiple-Category Graphs, IEEE Pacific Visualization Symposium, 121-128, 2009 .
- [2] D. Botstein, J. M. Cherry, M. Ashburner, C. A. Ball, J. A. Blake, H. Butler, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, Gene Ontology: tool for the unification of biology. Nat Genet 25: 25-29,2000.
- [3] iRefIndex, <http://irefindex.uio.no/wiki/iRefIndex>
- [4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation, Journal of Machine Learning Research, Vol. 3, pp. 993-1022, 2003.