Vol.2012-BIO-30 No.1
2012/8/9

平面上点集合の一様グリッドへの近似マッチング

下薗 真一[1,a)]　井上 健太郎[2]　倉田 博之[3,b)]

概要：代謝ネットワークのような規模が大きい相関関係に対して，われわれはインタラクティブに利用できる程度に高速で，かつすべてのノードラベルを重なることなく描画できるグリッドグラフ描画法を提案している [3], [4], [5]．その中でもハイブリッドレイアウトアルゴリズム [3] は，自由に選んだグラフ描画法を前処理として利用可能である．分割統治的し格子点への近似点集合照合を行うもので，CAD システムに必要な高速性と，グリッド配置の前後での変化を予測しやすくしている．しかしこの近似点集合の照合は，一般の点集合どうしの照合を対象としたものであるため，時間計算量と領域計算量の次数が高く，分割は定数サイズとする必要があった．本稿では，この平面上の近似点集合照合アルゴリズムを格子点への照合に最適化したアルゴリズムを示す．このアルゴリズムは，配置対象のグリッド領域の大きさにかかわらず，ノード数 n に関し $O(n^2)$ 時間で動作する．

キーワード：グラフ描画，代謝ネットワーク，近似幾何点集合照合

Efficient Grid Graph Layout by Approximate Point-set Matching on the Plane

Shinichi Shimozono[1,a)]　Kentaro Inoue[2]　Hiroyuki Kurata[3,b)]

Abstract: We present an $O(n^2)$-time grid-layout algorithm for graphs with n nodes that aligns every node on lattice exclusively. The grid graph layout approach for biochemical network maps [3], [4], [5] is a graph drawing scheme that place nodes on sparse grid points to prevent overlaps of node labels. Hybrid grid layout algorithm presented in [3] combines an arbitrary graph layout method, which may cause label overlaps, and the algorithm that transforms an arbitrary graph into a grid graph, by the divide-and-conquer and an approximate point-set matching on the plane. Although the computational results shown that the algorithm is time-efficient and achieves appropriate node layout, its approximate point-set matching requires more than $O(n^6)$ time and thus the minimization of transformations could be applied only on small areas. In this paper, we show this can be improved as $O(n^2)$ time algorithm, by utilizing the uniformness of grid point space.

Keywords: graph drawing, biochemical network, approximate point-set maching

1. はじめに

さまざまな相関関係のデータを俯瞰あるいは分析するうえで，グラフの自動描画 (e.g. [1]) は，情報処理システムが自動的な視覚化手段を提供するための方法の一つである．その描画においてグラフの辺は重視すべきだが，非常に規模が大きく未知の部分が多い代謝ネットワークの解析結果などを扱う場合は，ノードの配置や辺の交差などの細かい最適化の優先順位はあまり高くない．ユーザーインタラク

[1] 九州工業大学大学院大学院情報工学研究院知能情報工学研究系
Dept. of A.I., Kyushu Inst. of Tech., Kawazu 680-4, Iizuka, Japan
[2] 九州工業大学大学院情報工学研究院生命情報工学研究系
Dep. of Biosci. and Bioinfo., Kyushu Inst. of Tech., Kawazu 680-4, Iizuka, Japan
[3] 九州工業大学大学院情報工学研究院生命情報工学研究系
バイオメディカルインフォマティクス研究開発センター
Biomedical Informatics R&D Center, Kyushu Inst. of Tech., Kawazu 680-4, Iizuka, Japan
[a)] sin@ai.kyutech.ac.jp
[b)] kurata@bio.kyutech.ac.jp

ションを妨げない程度に高速で，ノードラベルのような文字情報をはっきり表示でき，ディスプレイ等の限りある面積で表示ができることがより強く求められる．

そのような場合に有効なネットワークの描画法として，我々は，一定の幅の格子点上にノードを配置し各ノードに必要な空間をあたえつつ，ノード間の相関関係は配置位置にゆるやかに反映されるグリッドグラフレイアウト法とそのアルゴリズムを提案し，実装すると同時に有効性を示してきた [3], [4], [5].

本稿では，特にハイブリッドアルゴリズム [3] において局所的にノードの格子点へのマッピングを求める部分アルゴリズム，二次元平面上の幾何点集合の近似マッチングアルゴリズムの高速化手法を提案する．

2. 記法と定義

ここでは議論を簡単にするため，ノードのラベルの大きさを考慮しない他のグラフ描画法によってノードの位置が与えられているか，もしくは，辺を考慮せずノードのラベルに必要な空間を確保するのが目的であると考える．つまり，ノードの格子点へのマッピングをグラフ描画とする．また，ノードのラベルはすべて，平面上で幅，高さともに d でおさまるものとする．

グラフ $G=(V,E)$ のノード数を $n=|V|$ とし，ノードは 1 から n の整数であると見なす．グラフ描画 D は，ノードの集合 $V=\{1,\ldots,n\}$ から平面上の座標 $\mathbb{Z}^{+2}$ への写像がグラフの描画である．特にグラフ描画がグリッド格子点 $L(d)=\{(kd,\ell d)\mid k,\ell\in\mathbb{Z}\}$ への写像 $D^{\dagger}:V\to L(d)$ であるとき，グリッドグラフレイアウトであるという．グリッドレイアウト問題は，入力としてグラフと任意のグラフ描画の組 (G,D) をうけとり，V から $L(d)$ への写像 $D^{\dagger}$ を求める計算問題である．

2.1 座標軸平行拡張

グラフ $G=(V,E)$ とグラフ描画 D から得られる平面上の点集合 S を $S=\{D(1),\ldots,D(n)\}$ とする．以下，議論を簡単にするため，$D(V)$ の任意の 2 点が X および Y の座標で同じ値をもたないものと仮定する．ある軸について同一の値を持つ，あるいは同一座標上にある複数の点の配置は，辺を考えない場合，一つ以上の格子点を占める大きさのある点として扱うことになり，単に最少領域への充填になるため，実装の議論で扱うことにする．

平面上の点集合 S の**境界矩形** *bounding rectangle* $B(S)$ とは，S に含まれる点の各座標値の最小値と最大値で定まる，辺が座標軸に平行ですべての点を含む最小の矩形である．点集合 S の**点** $p\in S$ **からの座標軸平行拡張列** *axis-parrallel expansion sequence* $S_1,\ldots,S_n=S$ は，$S_1=\{p\}$ からはじまり，要素数が 1 ずつ単調増加する S の部分集合の列で，その境界矩形が単調に拡張する列である; すなわち，$1\le i<n$ について $B(S_i)$ は常に $B(S_{i+1})$ に含まれる．

座標軸 X および Y にそった平面座標上の辞書式全順序 $\le_X$ と $\le_Y$ それぞれにおいて最も先となる S の点の組 $(p_{X,1},p_{Y,1})$ は，境界矩形 $B(S)$ の最左下点の座標をあたえる．これを $bl(S)$ と書くことにする．同様に，最右上点の座標を $tr(S)$ と書く．$B(S)$ の $bl(S)$ からの座標軸平行拡張列は，一点で $B(S)$ の左下角の座標をあたえる $p_{X,1}=p_{Y,1}$ があるか，もしくは 左下角を与える二点のうちいずれかを S_1, それにもう一方を追加して S_2 とする列で，単調に座標軸平行拡張されるものとする．

2.2 座標軸平行拡張による最適グリッドレイアウト

グリッドグラフレイアウト $D^{\dagger}$ は，$Grid_d:V\to\{0,\ldots,n-1\}^2$ であらわすことができる．すなわち $w=Grid_d(v)$ ならば $D^{\dagger}(v)=(w.x\cdot d,w.y\cdot d)$ である．あるグラフレイアウト D とその座標軸平行拡張 seq において，m 個目のノードまで間隔 d のグリッドグラフレイアウトであったとする．このとき，seq の $m+1$ 個目の点の位置を，m 個目までの点の境界矩形 $B(S_m)$ の外側で S_m の点から $tr(S_{m+1})$ の相対位置が格子点上となる位置へ移すのに必要な最少の移動量を $trans(B(S_m),S_{m+1})$ と書くことにする．

Definition 1. グラフ $G=(V,E)$ とその初期レイアウト $D:V\to\mathbb{Z}^2$ の組 (G,D) に対して，Seq を $S=D(V)$ の $bl(R(S))$ からの座標軸平行拡張とする．このとき，座標軸平行拡張 $seq=(S_1,\ldots,S_n)$ の D から $Grid_d$ における編集コスト $Cost_{D,d}(seq,n)$ は，帰納的に次のように定義する：

$$Cost_{D,d}(seq,n)=\begin{cases}0 & n=1,\\ Cost_{D,d}(seq,n-1)+ & \\ trans(B(S_{n-1}),tr(B(S_n))) & otherwise.\end{cases}$$

格子空間 $L(d)$ への座標軸平行拡張による最適グリッドレイアウト $Grid_d$ とは，(G,D) に対して，この編集コストが最少となる拡張 $seq=(S_1,\ldots,S_n)$ をあたえるグリッドレイアウトである．最適グリッドレイアウト問題は，あたえられた (G,D) に対して最適グリッドレイアウト，あるいはその左上角点列表現 $(tr(B(S_1)),\ldots,tr(B(S_n)))$ を求める計算問題である．

一般に，なんらかの編集コストを定義し，任意のグラフ描画をグリッド上に最少の編集コストで射影するという問題を考えた場合，その計算量的困難性などの結果は我々の知る限りない．グリッドではなく任意の点集合への写像とした場合，点集合間の編集コスト（距離）の最小化が NP 完全となる定義がある．一方で，射影を座標軸平行拡張に限れば，おなじ場合でも多項式時間でとくことができる．こ

の考えにもとづいて局所的なレイアウトを平面上の近似点集合マッチングで行ったのが文献 のアルゴリズムである．しかし，時間計算量が多項式とはいえ次数が高く，結果小さい領域にしか適用できなかった．このアルゴリズムを射影先が一様なグリッドであることを利用して高速化する．

3. 一様グリッドへの近似照合アルゴリズム

ある平面上の点集合 $S \subseteq \mathbb{Z}^2$ の部分集合 $S' \subseteq S$ が，構成途中の部分的なレイアウト関数によって，幅 d のグリッドレイアウトになっているとする．この S' の境界矩形 $B(S')$ の外側にありかつ右辺に最も近い点 $p \in S$, 外側で上辺に最も近い点を $q \in S$ とする．このとき，$B(S')$ に対し p を $trans(B(S'), p)$ だけ移動して格子点上に移し，拡張したものを右拡張 ${S'}_+$，q で同様に拡張したものを上拡張 ${S'}^+$ と書くことにする．この列は，座標軸平行拡張を与える．

一様グリッドへの近似照合では，この右拡張と左拡張のどちらが移動量の総和を小さくする拡張列となるかを，動的計画法によって求める．これにより，(G, D) から $Grid_d$ における編集コストを最少にする拡張列を求めることができる（証明は，文字列の Levenstein 距離を求めるアルゴリズム (e.g., [2] の場合と同様であるので省略する)．

DP 表を作るアルゴリズムの概略は，次のように書ける：

(1) $n \times n$ の表 $dist(n, n)$ を初期化する．
(2) $dist(1, 1) = trans(sort_X(1), sort_Y(1))$.
(3) X 座標優先でのソート列 $sort_X : \{1, \ldots, n\} \to V$ と Y 座標優先のソート列 $sort_Y : \{1, \ldots, n\} \to V$ を作成する．
(4) do
  (a) $ix := next_X(B(2, iy))$.
  (b) do
    (i) $iy := next_Y(B(ix, 2))$.
    (ii) $dist(ix, iy) = \max\{dist(ix, prev_Y(iy)) + trans(sort_Y[iy]), dist(prev_X(ix), iy) + trans(sort_X[ix])\}$
  (c) while $(iy \leq n)$.
(5) while $(ix \leq n)$.

ただし $B(j, k)$ は最上右角を $(sort_X(j), sort_Y(k))$ とする境界矩形であり，$next_X(B(j, k))$ は $B(j, k)$ に含まれない最も左（$sort_X$ で先）の点 $sort_X(next_X(B(j, k)))$ を参照する順位を返す関数である．これは，真偽値へのテーブルにより，$next_X$ および $next_Y$ の計算の際 j より右の点が $B(j, k)$ に含まれるかどうかを記録していく等で実装できる．$next_Y$ も同様である．$prev_X(B(j, k))$, $prev_Y(B(j, k))$ は逆に，最左および最上点を $B(j, k)$ に含まれる点集合から除去した点集合の最右および最上点である．

それぞれのループは，$next$ および $prev$ でスキップするが，スキップした回数と同じだけ $next$ および $prev$ 内でループを回るため，n 回の繰り返しと考えることができる．再帰式は定数時間で計算可能である．真偽値テーブルは，外側のループで毎回初期化する．それぞれ初期化に $O(n)$ 時間必要である．よって，DP 表の作成に必要な時間計算量は $O(n^2)$ である．

DP テーブル $dist$ を構築後，バックトラックによって実際のグリッドレイアウト $Grid_d$ を $O(n)$ 時間で求める．

以上から，一様グリッドへの近似アルゴリズムの時間計算量は，ノード数 n に関し $O(n^2)$ となる．

4. 同一座標値をもつ点への対応

これまでの議論では，すべての点について，他の点と同じ X もしくは Y 座標値はもたないという仮定をおいた．これは，同じ座標値をもつ点が複数あった場合，座標軸平行拡張，$trans$ 移動量の計算に組合せからの選択が生じるためである．実際の実装では，境界矩形の拡張が最も小さくなるように詰め込む，などタイブレークルールを導入し，$next$ および $prev$ の計算では同時に追加，削除を行う等の方法で解消する．

5. おわりに

本稿では，[3] のハイブリッドアルゴリズムにおける近似点集合照合アルゴリズムのオーダー $O(n^2m^4)$ （ただしグリッドの格子点数をノード数の二乗取った場合）を，グリッド格子点の一様性をもちいて $O(n^2)$ に抑える方法の概略を示した．オーダー評価の確認のために作成したＤＰテーブルにハッシュ表をつかった再帰法による簡易的な動的計画法のプログラムでも，この効果は確認できている．

今後は，移動量関数 $trans$ の，同一座標値をもつ点集合に対応し，かつ実際のレイアウトで効果的となる移動の設計などが課題となる．

参考文献

[1] Di Battista, G., Eades P., Tammasia, R. and Tollis, I. G., Algorithms for drawing graphs: an annotated bibliography, 入手先 ⟨http://www.cs.brown.edu/people/rt/gd-biblio.html⟩ (1994).
[2] Gusfield, D.: *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*, Cambridge University Press (1997).
[3] Inoue, K., Shimozono, S., Yoshida H., Kurata, H.: Application of approximate pattern matching in two dimensional spaces to grid layout for biochemical network maps, *PLoS one* 7(6): e37739, (2012). 入手先 ⟨http://dx.plos.org/10.1371/journal.pone.0037739⟩
[4] Kurata H, Inoue K, Maeda K, Masaki K, Shimokawa Y, et al.: Extended CADLIVE: a novel graphical notation for design of biochemical network maps and computational pathway analysis, *Nucleic Acids Res* 35: e134 (2007).
[5] Li, W., Kurata, H.: Visualizing global properties of large complex networks, *PLoS One* 3: e2541 (2008).