

階層構造を持つグラフの固有値を利用した発見手法

鈴木 純[†] 片山 薫[†]

[†] 首都大学東京システムデザイン研究科 〒191-0065 東京都日野市旭が丘 6-6

E-mail: [†]suzuki-jun@sd.tmu.ac.jp, ^{††}kaoru@comp.metro-u.ac.jp

あらまし たんぱく質構造やリレーショナルデータ等は、複雑な構造を持つ。これらは、グラフの各頂点が、違うグラフに関連付けられているような階層的な構造を持つグラフとして表現できる。このようなグラフを階層グラフと呼ぶこととし、グラフの場合と同様に階層部分グラフを定義する。部分グラフ同型性判定問題は NP 完全であり、階層部分グラフ同型性判定問題も多量の計算コストが必要になる。我々は階層グラフに含まれる階層部分グラフを固有値に関する Interlace 定理を用いることにより、効率的に発見する階層部分グラフ同型性判定の手法を提案する。本研究では、実験により組み合わせ的な手法のみで行うよりも、我々の手法が有用であることを確認した。

キーワード 部分グラフ同型性判定, Interlace 定理, 階層構造

Finding Hierarchical Subgraphs Efficiently by Eigenvalues

Jun SUZUKI[†] and Kaoru KATAYAMA[†]

[†] Graduate School of System Design, Tokyo Metropolitan University

Asahigaoka 6-6, Hino-shi, Tokyo, 191-0065 Japan

E-mail: [†]suzuki-jun@sd.tmu.ac.jp, ^{††}kaoru@comp.metro-u.ac.jp

1. はじめに

一方のグラフがもう一方のグラフに含まれるかどうかを判定する問題を部分グラフ同型性判定問題という。部分グラフ同型性判定問題は、画像における特徴量抽出 [13]、化合物のデータベース [17]、パターン認識などにおいて表れる。しかし、近年グラフは複雑なデータ構造を持ち、様々な情報々々関連づいたデータとして表現されることが多く、たんぱく質構造やネットワーク構造等は、従来のグラフでは十分な表現力を持たなかった。

これらは、グラフの各頂点が、違うグラフに関連付けられているような階層的な構造を持つグラフとして表現できる。このようなグラフを階層グラフと呼ぶこととし、グラフの場合と同様に階層部分グラフを定義する。部分グラフ同型性判定問題は NP 完全であり、階層部分グラフ同型性判定問題も多量の計算コストが必要になる。我々は階層グラフに含まれる階層部分グラフを固有値に関する Interlace 定理を用いることにより、効率的に発見する階層部分グラフ同型性判定の手法を提案する。この階層部分グラフ同型性問題は今後重要な課題になっていくと考えられる。

大規模なグラフを扱う場合、組み合わせ的な手法を使う前処理として数値計算的な手法は有効である [14]。グラフは対称行

列で表現することができ、行列とその部分行列の固有値には、Interlace 定理 [5] と呼ばれる関係があることが知られている。Interlace 定理はグラフの固有値に基づく手法であり、あるグラフの中に対象とするグラフが含まれないことを検出方法として用いることができる。グラフを隣接行列であらわすと、主部分行列は、誘導部分グラフを表しているため、Interlace 定理を利用することで誘導部分グラフではないものを見つけることができる。

2. 関連研究

部分グラフ同型判定問題に関する研究として、Ullmann [6] の手法や VF2 [7] がある。Ullmann の手法はバックトラックと呼ばれる手続きを用いて部分グラフ同型性判定を行う手法である。VF2 は 2 つのグラフの枝の間のマッピングの状態を更新していくことで 2 つのグラフの間の同型写像を作成する。VF2 は Ullmann の手法に比べて、使用するメモリ量が少ないという利点を持っている。

また、ある頂点にアイテム集合が付与されているグラフであるアイテム集合付きグラフ (IA グラフ) の解析 [16] や、多くの要素が複雑に絡み合うようなネットワーク構造のデータ解析 [18] などがある。これらのように、複雑な構造を持ったグラフに関する研究は限定的な分野ではあるが、されるようになって

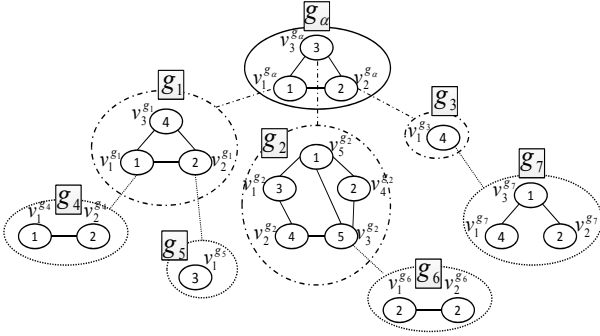


図1 例1の階層グラフ H

てきている .

3. 準備

3.1 階層部分グラフ同型

本研究では、無向グラフに焦点をあてる.

[定義 1] (グラフ) あるグラフを (V, E, L, μ) とする. $V = \{v_1, v_2, \dots, v_n\}$ を頂点の集合, $E = \{v_i v_j | v_i, v_j \in V, i \neq j\}$ を枝の集合, L を頂点と枝のラベルの集合, $\mu: V \cup E \rightarrow L$ を各頂点と各枝にラベルを割り当てる関数とする.

[定義 2] (階層グラフ) ある階層グラフ H を (G, ν) とする. G はグラフ集合 $\{g_i | i = \alpha, 1, 2, \dots, n\}$, グラフ g_α を階層グラフ H の一番高い階層のグラフだとする. $V(G)$ と $\nu(V)$ は, それぞれ $V(G) = \{v | v \in V(g), g \in G\}$ と $\nu(V) = \{g | g = \nu(v), v \in V\}$ を意味する. $\nu: V(G) \rightarrow G$ は頂点をグラフに対応させる関数とし, グラフ集合 G において, $\nu(\dots \nu(V(\nu(V(g_i)))) \cap \nu(\dots \nu(V(\nu(V(g_j)))) = \emptyset$ の条件を満たす.

[例 1] 階層グラフ H の例を図.1 に示す. この場合, グラフ集合は $G = \{g_\alpha, g_1, g_2, g_3, g_4, g_5, g_6, g_7\}$ となり, 一番高い階層のグラフは g_α である. $V(G)$ は全ての頂点の集合 $\{V(g_\alpha), V(g_1), \dots, V(g_7)\}$ を意味する. 枝ラベルには全て 1 を使い, グラフ g_α と g_2 に焦点をあてると, グラフ g_α はそれぞれ, $V(g_\alpha) = \{v_1^{g_\alpha}, v_2^{g_\alpha}, v_3^{g_\alpha}\}$, $E(g_\alpha) = \{v_1^{g_\alpha} v_2^{g_\alpha}, v_2^{g_\alpha} v_3^{g_\alpha}, v_3^{g_\alpha} v_1^{g_\alpha}\}$, $\mu(v_1^{g_\alpha}) = 1, \mu(v_2^{g_\alpha}) = 2, \mu(v_3^{g_\alpha}) = 3, \mu(E(g_\alpha)) = \{1, 1, 1\}$, $\nu(v_1^{g_\alpha}) = \{g_1\}$, $\nu(v_2^{g_\alpha}) = \{g_3\}$, $\nu(v_3^{g_\alpha}) = \{g_2\}$ となる. また, グラフ g_2 はそれぞれ, $V(g_2) = \{v_1^{g_2}, v_2^{g_2}, v_3^{g_2}, v_4^{g_2}, v_5^{g_2}\}$, $E(g_2) = \{v_1^{g_2} v_2^{g_2}, v_2^{g_2} v_3^{g_2}, v_3^{g_2} v_4^{g_2}, v_4^{g_2} v_5^{g_2}, v_5^{g_2} v_1^{g_2}, v_2^{g_2} v_3^{g_2}, v_3^{g_2} v_4^{g_2}, v_4^{g_2} v_5^{g_2}, v_5^{g_2} v_1^{g_2}\}$, $\mu(v_1^{g_2}) = 3, \mu(v_2^{g_2}) = 4, \mu(v_3^{g_2}) = 5, \mu(v_4^{g_2}) = 2, \mu(v_5^{g_2}) = 1, \mu(E(g_2)) = \{1, 1, 1, 1, 1, 1, 1, 1\}$, $\nu(v_1^{g_2}) = \emptyset, \nu(v_2^{g_2}) = \emptyset, \nu(v_3^{g_2}) = \{g_6\}, \nu(v_4^{g_2}) = \emptyset, \nu(v_5^{g_2}) = \emptyset$ となる.

[定義 3] (部分グラフ同型) $g^l = (V^l, E^l, L^l, \mu^l)$ を n 個の頂点を持つグラフ, $g^s = (V^s, E^s, L^s, \mu^s)$ を m 個の頂点を持つグラフとする ($m \leq n$).

部分グラフ同型性判定とは, 以下の条件を持つ単射写像 $f: V^l \rightarrow V^s$ が存在するか否かということである ($\forall (u, v) \in E^l$).

- (1) $(f(u), f(v)) \in E^s$.
- (2) $\mu^l(u) = \mu^s(f(u)), \mu^l(v) = \mu^s(f(v))$.
- (3) $\mu^l(u, v) = \mu^s(f(u), f(v))$.

これらの条件を満たすとき, g^s は g^l の部分グラフであり,

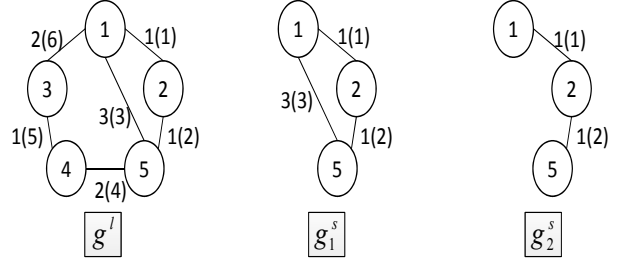


図 2 例: グラフ g^l , 誘導部分グラフ g_1^s , 部分グラフ g_2^s

$g^s \subseteq g^l$ と記述する.

[定義 4] (階層部分グラフ同型) $H^l = (G^l, \nu^l)$ と $H^s = (G^s, \nu^s)$ を階層グラフとし, それらの $G^l = \{g_i^l | i = \alpha, 1, 2, \dots, n\}$ と $G^s = \{g_j^s | j = \alpha, 1, 2, \dots, m\}$ をそれぞれのグラフ集合とする ($m \leq n$). 以下の条件を満たすとき, H^s は H^l の階層グラフとし, $H^s \subseteq H^l$ と記述する.

- (1) 全ての j において $g_j^s \subseteq g_i^l$ となる i が存在する.

グラフとは, 階層グラフの $\nu(V)$ が空集合である特別な場合と考えることができる. 以後, 話を明確にするために, 二つの階層を持つ場合に焦点をあてることにする.

3.2 グラフの対称行列表現

誘導部分グラフの表現方法

誘導部分グラフの場合, 以下のように隣接行列を用いて表現する. 隣接行列は対称行列であり, 行列の次数はグラフの頂点数に等しい.

[定義 5] グラフ $g = (V, E, L, \mu)$ を行列 A を用いて以下のように表現する. 対角成分には頂点のラベル, 非対角成分には枝のラベルを使用する. A_{ij} を行列 A の (i, j) 要素とする.

$$A_{ij} := \begin{cases} \mu(v_i) & (i = j) \\ \nu(v_i, v_j) & (i \neq j, (v_i, v_j) \in E) \\ 0 & (\text{otherwise}) \end{cases} \quad (1)$$

[例 2] 図.2 では, グラフ g_1^s はグラフ g^l の誘導部分グラフである. この場合, 以下のように g^l と g_1^s を行列表現する.

$$g^l = \begin{pmatrix} 1 & 1 & 3 & 0 & 2 \\ 1 & 2 & 1 & 0 & 0 \\ 3 & 1 & 5 & 2 & 0 \\ 0 & 0 & 2 & 4 & 1 \\ 2 & 0 & 0 & 1 & 3 \end{pmatrix}, g_1^s = \begin{pmatrix} 1 & 1 & 3 \\ 1 & 2 & 1 \\ 3 & 1 & 5 \end{pmatrix}$$

部分グラフの表現方法

誘導部分グラフではない部分グラフの場合, グラフを隣接行列で表現することができない.

[例 3] 図.2 のように, グラフ g_2^s は g^l の部分グラフではあるが, 誘導部分グラフではない場合, 隣接行列で表現するとそれぞれ以下ようになる.

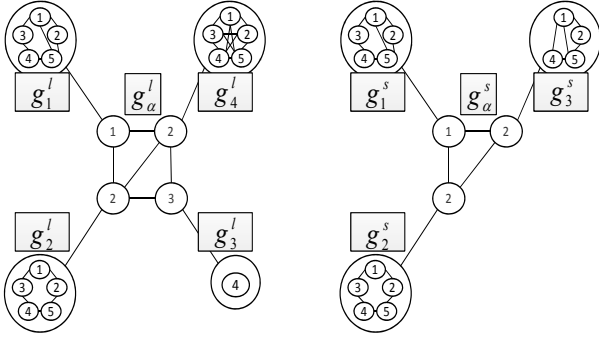


図3 例5: 階層グラフ H^l と階層グラフ H^s

$$g^l = \begin{pmatrix} 1 & 1 & 3 & 0 & 2 \\ 1 & 2 & 1 & 0 & 0 \\ 3 & 1 & 5 & 2 & 0 \\ 0 & 0 & 2 & 4 & 1 \\ 2 & 0 & 0 & 1 & 3 \end{pmatrix}, g_2^s = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 5 \end{pmatrix}$$

これは $S^T S = I$, $g_2^s = S^T g^l S$ となる $n \times m$ 行列 S が存在しない。

そこで, Haemers [5] が提案した接続行列を利用した表現方法を用いて, 以下のように表現する。

[定義6] グラフ $g = (V, E, L, \mu)$ を行列 M を用いて以下のように表現する。 M_{ij} を行列 M の (i, j) 要素とする,

$$M := \begin{pmatrix} P & N \\ N^t & Q \end{pmatrix}. \quad (2)$$

P, Q は, それぞれ以下の P_{ij}, Q_{ij} を要素とする対角行列である。

$$P_{ij} := \begin{cases} \mu(v_i) & (i = j) \\ 0 & (\text{otherwise}) \end{cases} \quad (3)$$

$$Q_{ij} := \begin{cases} \nu(v_k, v_l) & (i = j) \\ 0 & (\text{otherwise}) \end{cases} \quad (4)$$

[例4] 図.2 において, グラフ g_2^s は g^l の部分グラフである。このグラフ g_2^s の行列表現を以下に示す。

$$N^{g_2^s} : \begin{array}{c|cc} & (1) & (2) \\ \hline 1 & 1 & 0 \\ 2 & 1 & 1 \\ 5 & 0 & 1 \end{array}, g_2^s = \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 0 & \\ 0 & 2 & 0 & 1 & 1 & \\ \hline 0 & 0 & 5 & 0 & 1 & \\ 1 & 1 & 0 & 1 & 0 & \\ 0 & 1 & 1 & 0 & 1 & \end{array} \right)$$

階層グラフの行列表現

$H^l = (G^l, \nu^l)$ と $H^s = (G^s, \nu^s)$ を階層グラフとして, それぞれグラフ集合 G^l is $\{g_\alpha^l, g_1^l, \dots, g_n^l\}$ と G^s is $\{g_\alpha^s, g_1^s, \dots, g_m^s\}$ を持つとする ($m \leq n$)。

[例5] 図.3 における階層グラフ H^l を例とする。枝ラベルは全て1とすると, H^l の隣接行列表現は以下ようになる。

$$g_\alpha^l = \begin{pmatrix} 1 & 1 & 0 & 1 \\ 1 & 2 & 1 & 1 \\ 0 & 1 & 3 & 1 \\ 1 & 1 & 1 & 2 \end{pmatrix}, g_1^l = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 \\ 1 & 2 & 1 & 0 & 0 \\ 1 & 1 & 5 & 1 & 0 \\ 0 & 0 & 1 & 4 & 1 \\ 1 & 0 & 0 & 1 & 3 \end{pmatrix}, g_2^l = \begin{pmatrix} 4 \end{pmatrix},$$

$$g_3^l = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 & 1 \\ 1 & 1 & 5 & 1 & 1 \\ 1 & 1 & 1 & 4 & 1 \\ 1 & 1 & 1 & 1 & 3 \end{pmatrix}, g_4^l = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 \\ 1 & 2 & 1 & 0 & 0 \\ 0 & 1 & 5 & 1 & 0 \\ 0 & 0 & 1 & 4 & 1 \\ 1 & 0 & 0 & 1 & 3 \end{pmatrix}$$

グラフ $g_1^l, g_2^l, g_3^l, g_4^l$ は $\mu^l(g_\alpha^l)$ によって関連付けられたグラフである。

3.3 インタレース定理と部分グラフ

次の2つの実数列: $\alpha_1 \leq \alpha_2 \leq \dots \leq \alpha_n, \theta_1 \leq \theta_2 \leq \dots \leq \theta_m$ ($m \leq n$) において, 以下を満たすとき, 2つの実数列は Interlace するという。

$$\alpha_i \leq \theta_i \leq \alpha_{i+(n-m)} \quad (i = 1, \dots, m) \quad (5)$$

[定理1] Interlace 定理

S を $S^T S = I$ を満たす $n \times m$ の実行列とし, A を固有値 $\alpha_1 \leq \alpha_2 \leq \dots \leq \alpha_n$ を持つ $n \times n$ 対称行列とする。 $B = S^T A S$ を定義し, B は $\theta_1 \leq \theta_2 \leq \dots \leq \theta_m$ ($m \leq n$) の固有値を持つとする。このとき, B の固有値と A の固有値は Interlace する。

グラフ A, B を隣接行列で表現した場合, A, B の固有値が Interlace しなければ, グラフ B はグラフ A の誘導部分グラフではないといえる。本稿では, 固有値が Interlace しないことによって, 一方のグラフ g^s が, もう一方のグラフ g^l の誘導部分グラフではないことを検出する処理をフィルタリングと呼ぶ。

3.4 2分法に基づく固有値範囲比較

固有値の存在範囲によって Interlace 定理を利用するための, 計算手法について述べる [15]。固有値の大小関係を高速に比較するため, 本研究では行列の固有値の計算に2分法を用いる。2分法はシルベスターの慣性律を利用する。

n 次対称行列 A の固有値のうち, 負, 0, 正のもの個数, それぞれ $\tau(A), \xi(A), \pi(A)$ と定義する。

[定理2] シルベスターの慣性律

A が対称行列で, S が正則行列であるとし, $B = S^T A S$ とおく。そのとき, A と B は

$$\tau(A) = \tau(B), \xi(A) = \xi(B), \pi(A) = \pi(B)$$

という同じ慣性を持つ。

n 次対称行列 G^l と, m 次対称行列 G^s における2分法に基づく固有値比較のアルゴリズムを Algorithm1 に示す。固有値範囲更新の際は, 慣性に基づき, $g < z$ となる固有値の数を $Negcount(g, z)$ と定義する。 I を対角行列とすると, $Negcount(g, z)$ は行列 $g - zI$ の負の固有値の個数 $\tau(g - zI)$ を意味する。 g の小さい方から i 番目の固有値を λ_i とする。2分法は, λ_i が存在する上限 λ_i^{up} と下限 λ_i^{low} を初期値として与える。 $z = (\lambda_i^{up} + \lambda_i^{low})/2$ として, $Negcount(g, z)$ を計算する。 $Negcount(g, z)$ が i より小さい場合, λ_i^{low} の値を z で更新し, それ以外であれば λ_i^{up} の値を z で更新する。 $z = (\lambda_i^{up} + \lambda_i^{low})/2$ と, $Negcount(g, z)$ を使い, 上限と下限を繰り返し更新していく。固有値の存在範囲において, 固有値が Interlace するかどうかを調べることができた時点で終了する。

Algorithm 1 *Compare_Eigenvalues*

Inputs: g^l , (order n), g^s , (order m)

Outputs: g^s は g^l の部分グラフ候補である.

```

1:  $\{a_i^{low}\}, \{a_i^{up}\}, \{\theta_i^{low}\},$  and  $\{\theta_i^{up}\}$  を初期化する.
2: for  $i \leftarrow 1$  to  $m$  do
3:   while  $a_i^{up} \geq \theta_i^{low}$  do
4:     if  $a_i^{up} - a_i^{low} \geq \theta_i^{up} - \theta_i^{low}$  then
5:        $a_i^{up}$  or  $a_i^{low}$  の固有値範囲更新
6:     end if
7:     if  $a_i^{up} - a_i^{low} \leq \theta_i^{up} - \theta_i^{low}$  then
8:        $a_i^{up}$  or  $a_i^{low}$  の固有値範囲更新
9:     end if
10:    if  $\theta_i^{up} \leq a_i^{low}$  then
11:      return  $g^s$  は  $g^l$  の部分グラフではない.
12:    end if
13:  end while
14:  while  $\theta_i^{up} \geq a_{i+(n-m)}^{low}$  do
15:    if  $a_{i+(n-m)}^{up} - a_{i+(n-m)}^{low} \geq \theta_i^{up} - \theta_i^{low}$  then
16:       $\theta_i^{up}$  or  $\theta_i^{low}$  の固有値範囲更新
17:    end if
18:    if  $a_{i+(n-m)}^{up} - a_{i+(n-m)}^{low} \leq \theta_i^{up} - \theta_i^{low}$  then
19:       $\theta_i^{up}$  or  $\theta_i^{low}$  の固有値範囲更新
20:    end if
21:    if  $a_{i+(n-m)}^{up} \leq \theta_i^{low}$  then
22:      return  $g^s$  は  $g^l$  の部分グラフではない.
23:    end if
24:  end while
25: end for

```

4. 提案手法

4.1 階層部分グラフ同型性判定

本研究では階層グラフにおける、効率的な階層部分グラフ同型性判定, Algorithm2 を提案する. Algorithm2 は 5 つの *Step* で構成される.

Step1: g_α^s が g_α^l の部分グラフか否かを判定する (line 1 and 5). 初めに Interlace 定理による判定をし, 次に VF2 を利用する. g_α^s が g_α^l の部分グラフでない場合, この時点で H^s が H^l の階層部分グラフでない判定できる.

Step2: $V(g_\alpha^l)$ を $\nu^s(V(g_\alpha^l))$ によって頂点とグラフ G^l を対応づける (line 9). g_α^l が g_α^l の部分グラフであるので, $V(g_\alpha^s) \in V(g_\alpha^l)$ となる頂点が存在する. ゆえに, $V(g_\alpha^s) \in V(g_\alpha^l)$ となる頂点に関連付けられたグラフ $\nu^s(V(g_\alpha^l))$ のみを, $\nu^s(V(g_\alpha^s))$ と比較していけばよい. また, そのグラフ集合を $G^l = \{g^l | g_1^l, g_2^l, \dots, g_m^l\}$ と定義する.

Step3: 頂点と, その頂点に付いている枝のラベルでグラフ同型性問題の判定を行う (line 11 and 16). なぜなら g_α^s は複数のグラフ g^l の部分グラフ候補になりえるからである. グラフ同型性問題は, 全ての組み合わせで考えると莫大な時間を要す

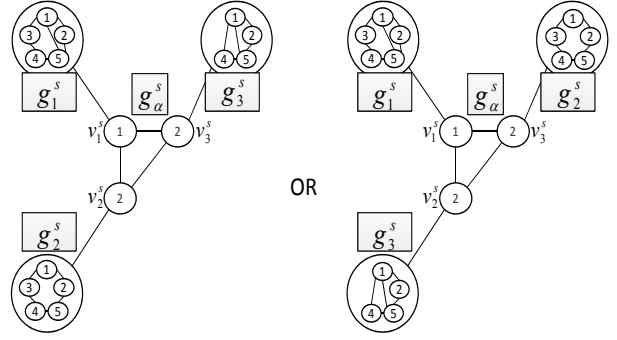


図 4 例 6 : 2通りの階層グラフ H^s

る為, 単純な処理として, 頂点とその頂点に付いている枝のラベルで比較するのみに留めておく. この処理の詳細は *Step* の後に記述する.

Step4: g_i^s が g_j^l の部分グラフでないものをフィルタリングする (line 14, 17, 19 and 23). g_i^s が, 全ての j における g_j^l の部分グラフではないと判定されたら, g_i^s は階層グラフ H^l に含まれていない為, この時点で H^s が H^l の階層部分グラフでない判定できる.

Step5: H^s が H^l の階層部分グラフか否かを判定する (line 30, 32 and 36). Interlace 定理によりフィルタリングされなかった部分グラフ候補を VF2 により部分グラフか否かを判定する. これは Interlace 定理によるフィルタリングでは候補を出力するのみで, 最終的には組み合わせの手法で確認しなければならない為である. そして最終的に H^s が階層グラフ H^l の階層部分グラフか否かを判定する.

前述と同条件の階層グラフ G_g^l と G_g^s が与えられたとき, グラフ同型性問題の判定をしないといけない. その例を以下に示す. [例 6] 図 4 のような形の階層グラフ H^s が考えられたとき, グラフ g_2^s とグラフ g_3^s が入れ替わった場合も, 階層グラフ H^s は同型である. ゆえに, $g_2^s \subseteq g_2^l$, $g_3^s \subseteq g_3^l$ となる可能性と, 同様に $g_3^s \subseteq g_2^l$, $g_2^s \subseteq g_3^l$ となる可能性がそれぞれある.

このように例 6 には二つの可能性がある. しかし, 同型判定を全ての組み合わせで完全に行うと, 処理時間が爆発的に増えてしまう可能性がある. そこで *Step3* のような単純な同型判定の処理を示す. 次の (1), (2) の条件を両方満たすとき, 同型の候補とする.

(1) 行列表現した際に, 頂点ラベルが同じ. (2) 頂点ラベルと枝ラベルを行で加算していった結果が同じ. この結果, 手順 4 における Interlace 定理による判定を行う回数は最悪で m^2 回である.

[例 7] 図 4 の g^s を隣接行列で表現したものと, その単純なグラフ同型性問題に対する処理を以下に示す.

$$g_\alpha^s = \begin{pmatrix} 1(v_1^s) & 1 & 1 \\ 1 & 2(v_2^s) & 1 \\ 1 & 1 & 2(v_3^s) \end{pmatrix}, g_\alpha^s = \begin{pmatrix} 1+1+1 \\ 1+2+1 \\ 1+1+2 \end{pmatrix} = \begin{pmatrix} 3 \\ 4 \\ 4 \end{pmatrix}$$

Algorithm 2 *Hierarchical_Subgraph_Isomorphism*

Inputs: $H^l = (G^l, \nu^l)$, $G^l = \{g^l | g^l_\alpha, g^l_1, \dots, g^l_n\}$, H^l の一番高い階層のグラフ g^l_α , $H^s = (G^s, \nu^s)$, $G^s = \{g^s | g^s_\alpha, g^s_1, \dots, g^s_m\}$, H^s の一番高い階層のグラフ g^s_α .

Outputs: H^s は H^l の階層部分グラフではない.

```
1: Compare_Eigenvalues ( $g^l_\alpha, g^s_\alpha$ )
2: if  $g^s_\alpha$  は  $g^l_\alpha$  の部分グラフではない. then
3:   return  $H^s$  は  $H^l$  の階層部分グラフではない.
4: end if
5:  $g^s_\alpha$  と  $g^l_\alpha$  の VF2 による部分グラフ同型性判定.
6: if  $g^s_\alpha$  は  $g^l_\alpha$  の部分グラフではない. then
7:   return  $H^s$  は  $H^l$  の階層部分グラフではない.
8: end if
9:  $G'^l \leftarrow \nu^s(V(g^l_\alpha))$ , このグラフ集合を  $G'^l = \{g'^l | g'^l_1, g'^l_2, \dots, g'^l_m\}$  とする.
10: for  $i \leftarrow 1$  to  $m$  do
11:    $A_i \leftarrow \mu(v^s_i) + \mu(v^s_i v^s_1) + \dots + \mu(v^s_i v^s_{i-1}) + \mu(v^s_i v^s_{i+1}) + \dots + \mu(v^s_i v^s_m)$ .
12: end for
13: for  $i \leftarrow 1$  to  $m$  do
14:    $G'^s_i \leftarrow \emptyset$  と初期化する.
15:   for  $j \leftarrow 1$  to  $m$  do
16:     if  $A_i = A_j$  then
17:       Compare_Eigenvalues ( $g'^l_j, g^s_i$ )
18:       if  $g^s_i$  は  $g'^l_j$  の部分グラフ候補である. then
19:          $G'^s_i \leftarrow G'^s_i \cup \{g^s_i\}$ 
20:       end if
21:     end if
22:   end for
23:   if  $G'^s_i = \emptyset$  then
24:     return  $H^s$  は  $H^l$  の階層部分グラフではない.
25:   end if
26: end for
27: for  $i \leftarrow 1$  to  $m$  do
28:   for  $j \leftarrow 1$  to  $m$  do
29:     if  $g'^l_j \in G'^s_i$  then
30:        $g^s_i$  と  $g^l_j$  の VF2 による部分グラフ同型性判定.
31:       if  $g^s_i$  は  $g^l_j$  の部分グラフではない. then
32:          $G'^s_i \leftarrow G'^s_i \setminus \{g^l_j\}$ 
33:       end if
34:     end if
35:   end for
36:   if  $G'^s_i = \emptyset$  then
37:     return  $H^s$  は  $H^l$  の階層部分グラフではない.
38:   end if
39: end for
```

(1) 頂点 v^s_2 と 頂点 v^s_3 のラベルは共に 2.

(2) 頂点 v^s_2 のラベルと、その頂点についている枝のラベルを足していくと $1 + 2 + 1 = 4$. 頂点 v^s_3 の場合も $1 + 1 + 2 = 4$ となり、結果が共に 4 である. このとき、 $\nu^s(v^s_3)$ に対応していた $\nu^s(v^l_\alpha)$ を、 $\nu^s(v^s_2)$ にも対応付ける. 同様に、 $\nu^s(v^s_2)$ に対応していた $\nu^s(v^l_\alpha)$ を、 $\nu^s(v^s_3)$ にも対応付ける.

本研究では、組み合わせアルゴリズムとして VF2 を使用する. Interlace 定理によりフィルタリングされないものについては、組み合わせ的なアルゴリズムを使用して部分グラフの判定を行う. 組み合わせ的なアルゴリズムと Interlace 定理を組み合わせることで、効率的に階層部分グラフを発見することができる.

5. 評価実験

全ての実験は CPU: Intel(R) Xeon (R) 2.50GHz, メモリ: 8GB, OS: Windows Vista Business の環境で行い、誘導部分グラフを利用して実験を行っている. 部分グラフの場合は、接続行列を利用して対称行列に変形することで、同様に使用でき、誘導部分グラフと同様の結果になるため特に記述はしない.

一つ目に、図.5 と 図.6 は、階層グラフ H^l のサイズの違いとして、グラフセットのグラフ数の多さと階層部分グラフ同型性判定に要する時間を示している. 二つ目に、表.1 から 表.4 は、グラフセットのグラフ数の数を固定し、階層グラフとして用いるグラフの条件を変えた場合の階層部分グラフ同型性判定に要する時間を示している.

本研究では我々の方法を評価する為に、VF2 のみで要した時間と、Interlace 定理によるフィルタリングと VF2 を組み合わせた手法に要した時間を比較している. 組み合わせアルゴリズムである VF2 [7] は部分グラフ同型性判定として高速であるとしられている. 実験では頂点と枝のラベルに正の整数を利用する.

5.1 階層部分グラフ同型性判定の処理時間

階層グラフの大きさが変化したとき、その階層部分グラフ同型性判定を行うまでの処理時間を示す. 図.5 では、階層グラフとして $H^l = (G^l, \nu^l)$, $H^s = (G^s, \nu^s)$ が、グラフ集合として $G^l = (g^l_\alpha, g^l_1, g^l_2, \dots, g^l_n)$, $G^s = (g^s_\alpha, g^s_1, g^s_2, \dots, g^s_m)$ がそれぞれ与えられたとする. グラフ g^l_α とグラフ g^s_α は、それぞれの階層グラフにおける一番高い階層のグラフとし、頂点数 n を持ち、25, 50, ... 200 と変化させていく.

頂点には 3 ラベル、枝には 1 ラベルを利用する. グラフ g^l_1, \dots, g^l_n には頂点数 50、頂点に 3 ラベル、枝数 350 前後、枝に 1 ラベルを利用する. グラフ g^s_1, \dots, g^s_m には頂点数 30、頂点に 3 ラベル、枝数 200 前後、枝に 1 ラベルを利用する. また、 $\nu(g^l_\alpha)$ で g^l_1, \dots, g^l_n が、 $\nu(g^s_\alpha)$ で g^s_1, \dots, g^s_m がそれぞれ関連付けられているとする. 図.6 では、 g^l_1, \dots, g^l_n と g^s_1, \dots, g^s_m の頂点ラベルを、2 ラベルに変化させた場合の処理時間である.

階層部分グラフ同型性判定の為に、Interlace 定理による処理を n 回 行い、 $n/2$ (切り上げ) 個のグラフのフィルタリングに成功し、残りの $n/2$ (切り捨て) 個のグラフを VF2 により判定した合計の処理時間を示す. その際 階層グラフの大きさであるグ

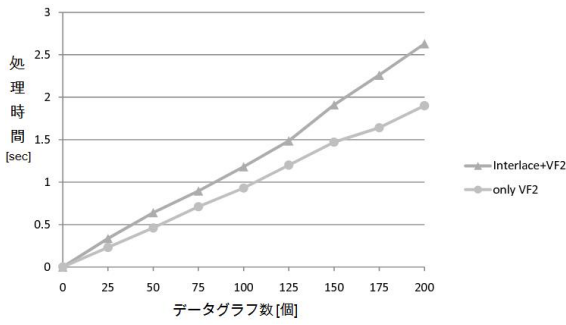


図 5 階層部分グラフ同型性判定の処理時間 (3 ラベル)

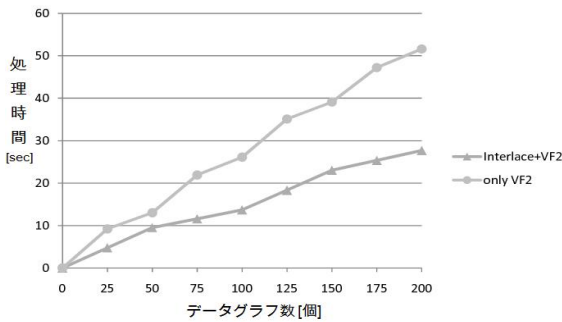


図 6 階層部分グラフ同型性判定の処理時間 (2 ラベル)

ラフ集合の数 n を 0 から 200 まで変化させていき、それぞれの処理時間を示す。

図.5 をみると、Interlace 定理と VF2 を利用した場合、VF2 のみの判定より処理時間がかかってしまっている。頂点ラベル数が 3 の場合には、VF2 が高速に行われるため、Interlace 定理を利用すると、結果的に処理時間が多くかかってしまう。

図 6 をみると、VF2 のみの判定よりも合計処理時間は短くなり、良い結果を示している。頂点ラベル数が 2 の場合には、VF2 による処理時間が多くかかるために、Interlace 定理によるフィルタリングによって、効率的に時間を短縮できることが確かめられた。

5.2 VF2 との処理時間比較

Interlace 定理によるフィルタリング性能は、

- (1) あるグラフと違うグラフの頂点数が近く、枝数が少ない。
- (2) あるグラフと違うグラフの頂点数の差が大きく、枝数が多い。

の二通りで有用であることが確かめられている [14].

階層グラフの頂点数と枝数を変えた時に、階層部分グラフ同型性判定までにかかる処理時間を表.1 から表.4 に示す。表の "VF2 only" は階層部分グラフ同型性判定に VF2 のみを使用した場合の時間、"Filtering" は Interlace 定理による処理に要する時間、"+VF2" は Interlace 定理によってフィルタリングされなかったグラフを VF2 により判定する時間、"Total" は "Filtering" と "+VF2" により要した階層部分グラフ同型性判定までにかかる時間をそれぞれ示している。要素における "-" は VF2 による処理時間が莫大過ぎて、時間が測れなかった際に使用している。

表 1 提案手法と VF2 の各処理時間 (g^l : 頂点数 100, 枝数 2000), (g^s : 頂点数 80), (ラベル数: 4)

枝数	Filtering [sec]	+VF2 [sec]	Total [sec]	VF2only [sec]
300	4.18	2089.01	2093.19	3678.30
400	4.22	897.07	901.29	1037.00
500	4.24	79.03	83.27	79.03
700	4.50	11.93	16.43	11.93
900	4.83	4.05	8.88	4.05
1100	5.17	1.39	6.56	1.39

表 2 提案手法と VF2 の各処理時間 (g^l : 頂点数 100, 枝数 2000), (g^s : 頂点数 80, 枝数 400)

ラベル数	Filtering [sec]	+VF2 [sec]	Total [sec]	VF2only [sec]
3	4.47	-	-	-
4	4.47	3358.00	3362.47	4206.06
5	4.48	1.26	5.74	1.764

表 3 提案手法と VF2 の各処理時間 (g^l : 頂点数 100, 枝数 2000), (g^s : 頂点数 80, 枝数 800)

ラベル数	Filtering [sec]	+VF2 [sec]	Total [sec]	VF2only [sec]
3	4.88	2606.47	2611.35	2606.47
4	4.84	8.38	13.22	68.38
5	4.82	0.46	5.28	0.46

実験は G^l を 100 個の頂点、2000 個の枝を持つグラフ 100 個のグラフ集合、 G^s を 100 個のグラフ集合として処理を行った。使用するグラフは、表.1 では、グラフ G_s の 50 個が G^l の誘導部分グラフであり、残りの 50 個が g^l の誘導部分グラフでないとする。表.2 から表.4 では、グラフ G^s 全てが G^l の誘導部分グラフでないとする。

表.1 は、 g^l と g^s の頂点と枝に使用するラベル数が 4、80 個の頂点を持つとし、枝数が 300, 400, ..., 1100 のグラフを使用した。表.2 は、 g^s は 80 個の頂点を持ち、400 個の枝を持つとし、 g^l と g^s の頂点と枝に使用するラベル数が 3, 4, 5 のグラフを使用した。表.3 と表.4 は、表.2 の、枝数を 800 と 1600 に変えたグラフを使用した。

表.1 において、枝数 300 のとき、提案手法は 2093.19[sec]、VF2 のみの手法は 3678.30[sec] となり、この場合、提案手法は有用である。しかしながら、多くの他の場合では VF2 のみで行う手法の法が提案手法よりも早い結果となっている。

また表.2 から表.4 では、 g^l と g^s 使用ラベル数を 3 から 5 に変化させている。ここから分かるように、VF2 の処理時間はラベル数にかなり依存し、枝が少ない場合は組み合わせ爆発を起こすので顕著に時間を多く要する。よって、表.4 のように枝数が多い場合、多少フィルタリングされたとしても有用ではない。しかし、表.2 や表.3 のように、あるグラフと違うグラフの頂点数が近く、枝数が少ない場合、提案手法は有用である。

表 4 提案手法と VF2 の各処理時間 (g^l : 頂点数 100, 枝数 2000), (g^s : 頂点数 80, 枝数 1600)

ラベル数	Filtering [sec]	+VF2 [sec]	Total [sec]	VF2only [sec]
3	5.83	0.15	5.88	0.35
4	5.82	0.024	5.844	0.042
5	5.78	0.015	5.795	0.026

6. 結 び

本稿では, Interlace 定理による効率的な階層部分グラフの発見手法を提案した. 実験から, Interlace 定理を利用することによって, 組み合わせアルゴリズムのみでは時間がかかるものに対して, 高速に階層部分グラフを発見できることが分かった. 近年の複雑な情報が関連づいたデータにおいて, 本提案は様々な分野で利用できるものであろう.

現状の問題点として提案手法には, 階層グラフであるグラフがより大きな階層を持っていた場合, 計算量が爆発的に増える可能性がある. よって今後の課題として, (1) より複雑な階層を持っている階層グラフへの処理方法. (2) Interlace 定理によるフィルタリングの性能の向上. (3) 固有値の計算速度の向上. 等があり, テンソルデータを利用した部分グラフ同型性判定が, より高速にフィルタリング処理ができ, かつ複雑な構造に対応させることができると考えられる.

文 献

- [1] Horst Bunke. Graph matching: Theoretical foundations, algorithms, and applications. *Vision Interface 2000*, pp. 82–88, May 2000.
- [2] X. Yan C. Chen, P.S. Yu, J.Han, D.-Q. Zhang, and X. Gu. Towards graph containment search and indexing. *VLDB'07*, pp. 926–937, 2007.
- [3] J. Cheng, Y. Ke, W. Ng, and A. Lu. Fg-index: towards verification-free query processing on graph databases. *SIGMOD*, pp. 857–872, 2007.
- [4] J. W. Demmel. Applied numerical linear algebra. *Society for Industrial and Applied Mathematics*, 1997.
- [5] Willen H. Haemers. Interlacing eigenvalues and graphs. *Linear Algebra and its Applications*, Vol. 226–228, pp. 593–616, 1995.
- [6] J.R. Ullmann. An algorithm for subgraph isomorphism. *ACM*, Vol. 23, No. 1, pp. 31–42, January 1976.
- [7] L.P. Cordella, P. Foggia, C. Sansone, and M. Vento. An improved algorithm for matching large graphs. *the 3rd IAPR TC-15 Workshop on Graph-based Representations in Pattern Recognition*, pp. 149–159, May 2001.
- [8] B. Luo and E. R. Hancock. Structural graph matching using the em algorithm and singular value decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10), pp. 1120–1136, 2001.
- [9] B. D. McKay. Practical graph isomorphism. *Congressus Numerantium*, Vol. 30, pp. 45–87, 1981.

- [10] S. Nijssen and J. N. Kok. A quickstart in frequent structure mining can make a difference. *KDD*, pp. 647–652, 2004.
- [11] J. T. L. W. D. Shasha and R. Giugno. Algorithmics and applications of tree and graph searching. *In Symposium on Principles of Database Systems*, pp. 39–52, 2002.
- [12] S. Zampelli, Y. Deville, and P. Dupont. Declarative approximate graph matching using a constraint approach. *Second International Workshop on Constraint Propagation and Implementation*, Vol. 1, , October 2005.
- [13] 諏訪美佐子, 藤本克仁. グラフ表現を利用した手書き数字列のセグメンテーション方式. *Technical report of IEICE. PRMU*, Vol. 107, No. 491, pp. 161–166, February 2008.
- [14] 長屋未来, 片山薫. 大規模グラフを対象とした部分グラフ同型判定における interlace 定理の利用. *DEWS*, 2006.
- [15] 長屋未来, 片山薫. 部分グラフ同型判定のための 2 分法を利用した固有値比較手法の提案. *DEWS*, 2008.
- [16] 山田恵, 瀬々潤. アイテム集合付き部分グラフに対する事後処理法の提案. *DEIM*, 2010.
- [17] 和田貴久, 大野博之, 稲積宏誠. 部分構造に基づく構造類似性を用いた特徴抽出システムとその応用. *The Database Society of Japan*, Vol. 7, No. 1, pp. 187–192, June 2008.
- [18] 鹿島久嗣. ネットワーク構造予測 (特集ベジアンネットワークと確率的情報処理の新展開). *人工知能学会誌*, Vol. 22, No. 3, pp. 344–351, 2007-05-01.