D-020

固有値を用いたグラフ検索のためのグラフの行列表現の比較

Experimental Comparison of Matrix Expressions for Finding Graphs by Eigenvalues

佐藤　正史† 片山　薫‡
Masashi Sato Kaoru Katayama

1 はじめに

近年，図形やたんぱく質，DNA，化合物などがグラフで表現されており，グラフは複雑な構造を表すための重要なデータモデルとなっている．グラフを行列で表現し，固有値を求めて Interlace 定理を用いることによって，部分グラフでないものを見つけることができる．本稿では，グラフの行列表現によって部分グラフではないものを発見する精度 (フィルタリング精度) や処理時間がどのように変化するかについて実験的に比較した．

2 関連研究

部分グラフ同型性判定に用いる行列表現に Haemers[1] が提案した行列表現がある．長屋ら [2] が提案する行列表現はこれをベースにしている．Nguyen ら [3] はグラフをクラスタリングするためスペクトルを最適化するアルゴリズム OPTSPEC を提案した．これにより，多くの場合において Interlace 定理による誤判定数を大幅に減少できることを示した．

3 比較するグラフの行列表現

本稿では，ラベル付き無向グラフ $G = (V, E, L_V, L_E, \mu, \nu)$ を対象とする．$V = \{v_1, v_2, \ldots, v_{|V|}\}$ を頂点の集合，$E = \{e_1, e_2, \ldots, e_{|E|}\} = \{v_i v_j | v_i, v_j \in V, i \neq j\}$ を枝の集合，L_v を頂点ラベルの集合，L_E を枝のラベルの集合とする．μ, ν をラベルを与えるための関数とする．$\mu : V \to L_V, \nu : E \to L_E$ は，個々の頂点と枝にラベルを割り当てる関数である．本稿では下記の行列表現 A, B, C，及び頂点と枝のラベルを全て 1 にした行列表現 A', B', C' を比較する．

$$A = \begin{bmatrix} P & N \\ N^t & Q \end{bmatrix} \tag{1}$$

$$B = (b_{ij}) = \begin{cases} \nu(e_i) & (i = j) \\ \mu(v_k) & (e_i = v_k v_l,\ e_j = v_k v_m \in E) \\ 0 & (\text{それ以外}) \end{cases} \tag{2}$$

$$C = N^t N \tag{3}$$

ここで,

$$N = (n_{ij}) = \begin{cases} \nu(e_i) & (v_i v_k = e_j \in E) \\ 0 & (\text{それ以外}) \end{cases} \tag{4}$$

$$P = (p_{ij}) = \begin{cases} \mu(v_i) & (i = j) \\ 0 & (\text{それ以外}) \end{cases} \tag{5}$$

$$Q = (q_{ij}) = \begin{cases} \nu(e_i) & (i = j) \\ 0 & (\text{それ以外}) \end{cases} \tag{6}$$

である。

4 グラフと固有値

$\alpha_1 \le \alpha_2 \le \cdots \le \alpha_n$と$\beta_1 \le \beta_2 \le \cdots \le \beta_m (m < n)$ の 2 つの数列が以下の式を満たすとき，$\{\beta_i\}$ は $\{\alpha_i\}$ を Interlace すると言う．

$$\alpha_i \le \beta_i \le \alpha_{i+(n-m)} \tag{7}$$

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$\alpha_1 \le \alpha_2 \le \ldots \le \alpha_n$の固有値を持つ $n \times n$ の対称行列とする．$B = S^T AS$ を定義し，B は$\beta_1 \le \beta_2 \le \cdots \le \beta_m$($m < n$) の固有値を持つとする．このとき，B の固有値は A の固有値をInterlace する．

グラフ G を 3 章の行列表現では表したものを M とすると，その部分グラフは $S^T S = I$ となる S を用いて $S^T MS$ と表すことができる．

5 実験的比較

行列表現 A, B, C, A', B', C' のフィルタリング精度と処理時間を比較する．グラフのデータはグラフジェネレーターで生成される人工グラフデータを用いた．グラフが 1000 個あるデータセットを 2 つ用意する．2 つのデータセットは部分グラフにならない関係であり，本実験では Interlace しないものをどれだけ出せたかでフィルタリング精度を図る．コンピュータのスペックは OS は Windows7，CPU は Intel(R) Core(TM) i5 CPU，メモリは 8.00GB である．

実験 1

頂点 35 個枝 297 個と頂点 10 個枝 23 個のグラフを 1000 組判定する. その後頂点を 15,20,25,30 と変えていき行う. 頂点と枝に割り当てられるラベル数は 1〜3 である．各頂点の枝の数は頂点に対して付けることができる数の約 50 %の値である. 結果を表 1，表 2 に示す.

行列表現 A' は他の行列表現と比べても最も優れた結果であった．しかし，A と A' は 2 つのグラフの枝数の差が大きい場合，1 つも正確な結果を出せないことが解かった．行列表現 B は 2 つのグラフの枝数の差が小さい場合，フィルタリング精度は最も優れた結果が出た．B' の処理時間は他の行列表現と比べても最も速かった．表 1，表 2 の結果から最もフィルタ

† 首都大学東京大学院システムデザイン研究科
‡ 首都大学東京大学院

頂点数	10	15	20	25	30
A	0	0	21	104	181
A'	0	**395**	**551**	**530**	594
B	11	60	148	349	**770**
B'	**76**	139	135	106	154
C	61	133	136	107	148
C'	31	101	67	74	210

表 1　発見された部分グラフでないグラフの数

頂点数	10	15	20	25	30
A	548.6	560.4	581.6	643.1	792.4
A'	215.2	215.9	224.1	252.3	324.6
B	422.1	425.2	438.3	481.6	601.2
B'	**111.0**	**113.9**	**122.2**	**126.2**	**163.0**
C	133.6	132.6	139.6	157.0	193.2
C'	134.4	133.7	142.0	160.1	195.6

表 2　処理時間 [sec]

リング精度が高いのが行列表現 A，最も処理時間が速いのが行列表現 B' である．

実験 2

頂点 20 個枝 187 個と頂点 21 個枝 207 個のグラフを 1000 組判定する．その後頂点を 22,23,24,25 と変えていき行う．各頂点の枝の数は頂点に対して付けることができる数の約 97 %の値である．頂点と枝に割り当てられるラベル数は 1〜3 である．結果を表 3，表 4 に示す．

頂点数	21	22	23	24	25
A	836	427	178	80	59
A'	874	**947**	**895**	**940**	**953**
B	**990**	908	815	735	627
B'	313	237	325	120	49
C	439	159	56	52	53
C'	314	237	324	120	49

表 3　発見された部分グラフでないグラフの数

頂点数	21	22	23	24	25
A	330.8	371.6	437.5	539.0	636.9
A	110.1	126.7	143.0	162.7	184.2
B	255.0	301.7	358.7	427.3	511.6
B'	**71.0**	**79.8**	**90.9**	**105.5**	**120.6**
C	87.5	97.8	107.3	122.6	141.7
C'	77.9	88.5	99.3	114.6	132.0

表 4　処理時間 [sec]

フィルタリング精度と処理時間が最も良かった行列表現，基本的な結果は最初の実験と同じであった．行列表現 A' と B は 2 つのグラフの枝数の差が小さい場合，80 %〜90 %のフィルタリング結果を出すことができることが解かった．

実験 3

実験 2 の頂点と枝に割り当てられるラベル数を 1〜7 に変えて行う．結果を表 5，表 6 に示す．

頂点数	21	22	23	24	25
A	935	614	304	140	89
A'	854	948	**880**	**916**	**940**
B	**998**	**960**	872	769	658
B'	320	238	315	125	72
C	498	256	107	75	73
C'	320	238	314	126	72

表 5　発見された部分グラフでないグラフの数

頂点数	21	22	23	24	25
A	386.5	450.0	527.3	626.6	735.2
A'	112.0	127.6	143.8	166.8	185.4
B	299.3	352.4	418.1	504.2	607.2
B'	**75.0**	**84.4**	**96.4**	**110.5**	**124.6**
C	86.6	98.3	115.1	130.0	150.9
C'	82.7	93.8	106.9	120.8	139.0

表 6　処理時間 [sec]

実験 2 と比べてフィルタリング精度が良くなった．ラベル数が多いとフィルタリング精度が上がることが解かった．

6　おわりに

本研究では部分グラフ同型性判定に用いるグラフの行列表現の比較実験を行い，有効なケースを確認した．行列表現 A' と B は 2 つの枝数の差が小さい場合，フィルタリング精度が 80 %〜90 %である．今後の課題として，フィルタリング精度と処理時間が両方優れている行列表現を発見することが求められる．

参考文献

[1] Willem H.Haemers. Interlacing Eigenvalues and Graphs. linear Algebra Appl. 226, pp.593-616. 1995.
[2] 長屋未来，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. 第 19 回データ工学ワークショップ DEWS 2008，2008.
[3] Nguyen Duy Vinh，Akihiro Inokuchi，Takashi Washio. Graph Classification Based on Optimizing Graph Spectra. Discovery Science，pp.205-220，2010.