3D2-4

グラフスペクトルに基づいた大規模頻出部分グラフマイニング

A study on large graph mining

グェン　ズィ　ヴィン　　大原　剛三　　鷲尾　隆
Duy Vinh Nguyen　　Kouzou Ohara　　Takashi Washio

大阪大学 産業科学研究所
The Institute of Scientific and Industrial Research (ISIR), Osaka University

Most of the existing frequent subgraph mining algorithms do not have good performance to find large frequent subgraphs. In this study, we assess some approach of large graph mining.

1. はじめに

近年，化学，コンピュータネットワークなどの分野をはじめ，グラフとして表現したデータを直接扱うことが急速に増加しつつある．そのため，与えられた単一のグラフ，もしくは複数のグラフから特徴的な部分グラフを発見するグラフマイニングが盛んに研究されており，その中でも，あるグラフデータベースに頻出する部分グラフを求める頻出部分グラフマイニングは主要な課題となっている．これまでに，頻出誘導部分グラフを完全に検索する AGM[Inokuchi 00]，一般連結頻出誘導部分グラフを完全検索する AcGM[Inokuchi 02]，頻出連結部分グラフを完全検索するアルゴリズム gSpan [Yan 02]，Gaston [Nijssen 04], FSG [Kuramochi 01] などが提案されている．しかしながら，これらの従来のアルゴリズムでは，大規模な頻出一般連結部分グラフを実用時間内に求めることは困難であった．なぜなら，それらは大規模頻出一般連結部分グラフを求めるために，それより小さい一般連結部分グラフを全て列挙する必要があるからである．そこで，本研究では，グラフスペクトルを用いて，従来の手法が現実的時間でマイニングできない大きさをもつ大規模な頻出一般連結部分グラフを，高速にマイニングする大規模グラフマイニング手法を提案する．

2. 関連研究

グラフスペクトルとは，グラフ G を表現する対称行列の固有値からなるベクトルである．一方，Interlace 定理は，ある対称行列 H の固有値とその主小行列 h の固有値の関係を示す．

定理 1:(Interlace 定理)[Ikebe 87]
ある $n \times n$ 対称行列 H の固有値 $\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$ と，その $m \times m$ の主小行列 h の固有値 $\theta_1 \leq \theta_2 \leq ... \leq \theta_m (m < n)$ が与えられたとき，次の関係が成り立つ．

$$\lambda_k \leq \theta_k \leq \lambda_{k+(n-m)} \tag{1}$$

グラフ G を対称行列である隣接行列 $A(G)$ で表現した場合，その誘導部分グラフは $A(G)$ の主小行列となるため，G の誘導部分グラフのグラフスペクトルが取るべき値の範囲を Interlace 定理により求めることができ，それが G の誘導部分グラフとなる為の必要条件となる [Haemers 95]．一方，G の接続行列 $N(G)$ は，G の頂点数を x，辺数を y としたとき，$x \times y$ 行列となることから一般には対称行列とはならない．しかし，以下のような拡張接続行列 $E(G)$ を考えた場合，$E(G)$ は対称行列となるため Interlace 定理を適用でき [Haemers 95]，同様に，あるグラフが G の一般部分グラフとなるための必要条件を求めることができる．

$$E(G) = \begin{pmatrix} 0 & N(G) \\ N(G)^T & 0 \end{pmatrix}. \tag{2}$$

いま，正方行列 M の次数を $dim(M)$ と表すと，$dim(E(G)) = x + y$ となる．このとき，ある連結グラフ G が一般部分グラフ g をもつための必要条件は定理 1 より導かれる以下の系により与えられる．

系 1：
ある連結グラフ G が一般部分グラフ g をもつための必要条件は，$E(G)$ の固有値を $\lambda_1 \leq \cdots \leq \lambda_{dim(E(G))}$，$E(g)$ の固有値を $\theta_1 \leq \cdots \leq \theta_{dim(E(g))}$ としたとき，$k = 1, ..., dim(E(g))$ に対して，

$$\theta_k \in int_k(G, g) = [\lambda_k, \lambda_{k+dim(E(G))-dim(E(g))}] \tag{3}$$

が成り立つことである．□

3. 提案原理

本研究で提案する大規模グラフマイニング手法は，データグラフ集合 D，最小支持度 $minsup$，最小グラフサイズ $minsize$ が与えられたときに，$minsize$ 以上で，少なくとも $minsup$ 個の D 中のデータグラフに現れる一般連結部分グラフ（頻出一般連結部分グラフ）を求める．そのために本章では，まず，Interlace 定理に基づいて導かれる以下の補題を示す．

補題 1:
ある連結グラフ G の頂点数 m の任意の部分木 g_{tree} に関して，Interlace 定理により与えられる $E(g_{tree})$ の固有値の取り得る区間 $int_k(G, g_{tree})(k = 1, ..., 2m-1)$ は，他のいかなる頂点数 $x(\geq m)$ の一般連結部分グラフ g に関する区間 $int_k(G, g)$ をも包含する．

連絡先: NGUYEN DUY VINH
〒 567-0047 大阪府茨木市美穂ヶ丘 8-1
大阪大学 産業科学研究所
duyvinh@ar.sanken.osaka-u.ac.jp

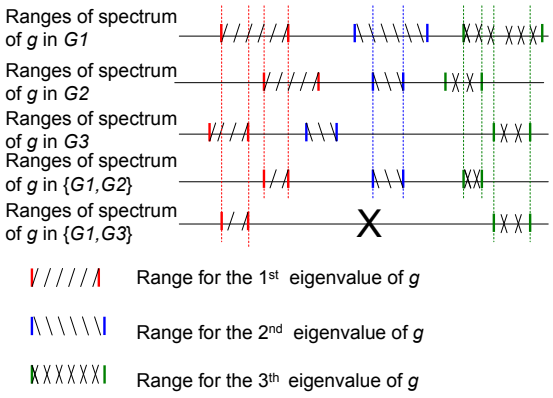

図 1: 共通一般連結部分グラフの取り得る固有値の範囲の計算

証明:
頂点数 m の部分木 g_{tree} の辺数は $m-1$ なので，定義より
$dim(E(g_{tree})) = 2m-1$
である。また，同じく g の辺数を $e(g)$ とすると
$e(g) \geq x-1$
であるので，$x \geq m$ より
$dim(E(g)) = x + e(g) \geq 2x-1 \geq dim(E(g_{tree}))$ (*)
である．一方，系 1 より
$int_k(G,g) = [\lambda_k, \lambda_{k+dim(E(G))-dim(E(g))}]$
$(k = 1, ..., dim(E(g)))$
$int_k(G,g_{tree}) = [\lambda_k, \lambda_{k+dim(E(G))-dim(E(g_{tree}))}]$
$(k = 1, ..., dim(E(g_{tree})))$
であるので，式 (*) より
$int_k(G,g) \subseteq int_k(G,g_{tree})$
$(k = 1, ..., dim(E(g_{tree})) = 2m-1)$
である．□

データグラフ集合 D の任意の頻出一般連結部分グラフは，$minsup$ 個以上のデータグラフ集合 $D_s = G_1, ..., G_n (\subseteq D)$ に共通に含まれる一般連結部分グラフである．このことから，D_s が共通一般連結部分グラフをもつためには，Interlace 定理に基づく各 G_i の一般連結部分グラフの第 k 固有値が取り得る範囲が共通領域をもたなければならい．たとえば図 1 では，ある一般連結部分グラフ g がデータグラフ $G1$, $G2$ それぞれの一般連結部分グラフとなるためのグラフスペクトルの取り得る値の範囲の共通領域が存在するため，$\{G1, G2\}$ が共通一般連結部分グラフをもつ可能性があると判断でき，その共通領域が $\{G1, G2\}$ の共通一般連結部分グラフ g のグラフスペクトルの取り得る範囲となる．一方，$\{G1, G3\}$ に関しては，その第 2 固有値について共通一般連結部分グラフのグラフスペクトルが取り得る領域が存在しないため，$\{G1, G3\}$ は共通一般連結部分グラフを持ち得ないと判定できる．以上のような各固有値が取り得る範囲の共通領域について，以下の補題が成り立つ．

補題 2:
D の任意の部分集合 $D_s = \{G_1, ..., G_n\}$ が大きさ $minsize$ 以上の何らかの共通一般連結部分グラフをもつ必要条件は，すべての $k = 1, ..., 2minsize-1$ について，
$\max_{i=1,...,n}(\lambda_k^{G_i}) \leq \min_{i=1,...,n}(\lambda_{k+dim(E(G_i))-2minsize+1}^{G_i})$
が成り立つことである．ここで，$\lambda_k^{G_i} (k = 1, ..., dim(E(G_i)))$ は，データグラフ $G_i \in D_s$ の第 k 固有値である．

証明:
各 $G_i \in D_s$ がある共通一般連結部分グラフ g をもつためには，系 1 より $E(g)$ の $(k = 1, ..., dim(E(g)))$ に関するすべての固有値 θ_k とすべての $G_i \in D_s$ について，
$\theta_k \in int_k(G_i, g)$
が成立しなければならない．従って，すべての $(k = 1, ..., dim(E(g)))$ について
$INT_k(D_s, g) = \cap_{i=1,...,n} int_k(G_i, g) \neq \phi$ (**)
でなければならない．
補題 1 より頂点数 $minsize$ の任意の部分木 g_{tree} に関して，
$int_k(G,g) \subseteq int_k(G, g_{tree}) (k = 1, ..., 2minsize-1)$
であるので，いずれかの $k = 1, ..., dim(E(g_{tree}))$ について
$INT_k(D_s, g_{tree}) = \cap_{i=1,...,n} int_k(G_i, g_{tree})$
$= [\max_{i=1,...,n}(\lambda_k^{G_i}), \min_{i=1,...,n}(\lambda_{k+dim(E(G_i))-2minsize+1}^{G_i})] = \phi$
すなわち，
$\max_{i=1,...,n}(\lambda_k^{G_i}) > \min_{i=1,...,n}(\lambda_{k+dim(E(G_i))-2minsize+1}^{G_i})$
であれば，式 (**) は成立せず，$D_s = \{G_1, ..., G_n\}$ は大きさ $minsize$ 以上のいかなる共通一般連結部分グラフをもたない．従って，この対偶命題である本補題は成立する．□

4. 提案手法

4.1 提案手法の概要

従来の多くのグラフマイニング手法では，各データグラフの部分グラフを候補頻出部分グラフとして列挙し，その出現頻度（候補頻出部分グラフを含むデータグラフ数）を計算して，$minsup$ 以上の頻出部分グラフを導出する [Inokuchi 00][Yan 02]．しかし，大きさ n のデータグラフの候補頻出部分グラフ数は最大 2^{n^2} 個であるため，この方法は n が大きいと計算コストが非常に高くなる可能性があり，大規模頻出部分グラフのマイニングには適さない．そのため，本稿で提案する大規模グラフマイニング手法では，グラフデータベース D の各データグラフから候補頻出部分グラフを列挙するのではなく，データグラフの組合せ $D_s \subseteq D$ を列挙し，D_s 内のすべてのグラフに共通する一般連結部分グラフを探索する．あるデータグラフの組合せ D_{si} の共通一般連結部分グラフの集合 $\mathbf{g}$ が求められたとき，D_{si} を包含するデータグラフの集合 D_{sj} の共通一般連結部分グラフの集合は $\mathbf{g}$ の部分集合となる．そのため，$minsup$ 個のデータグラフの組み合わせの共通一般連結部分グラフを求めることで，全ての頻出一般連結部分グラフを求めることができる．そこで，$n = 1$ からはじめて，共通一般部分グラフの存在する可能性がある限り，$n = minsup$ に至るまで D_s の列挙を行う．このため，可能なデータグラフ組み合わせの列挙数は最大で $\sum_{n=1}^{minsup} {}_{|D|}C_n$ 個だけある．

列挙した各データグラフの組合せ D_s が，共通一般連結部分グラフを持つか否かを判定する為に，提案する大規模グラフマイニング手法では Interlace 定理を用いる．実際には，D_s に対して，補題 3 を用いて，大きさ $minsize$ の共通一般連結部分グラフが存在し得るかどうかを確認すれば，補題 4 より大きさ $minsize$ 以上の共通一般連結部分グラフが存在し得るかどうかの判定には十分である．

これらのことを用いて，提案する大規模グラフマイニング手法では，図 2 に示す列挙木を深さ優先探索でたどりつつデータグラフの組合せを列挙し，$minsize$ の共通一般連結部分グラフを持つ必要条件を満たす $minsup$ 個のデータグラフからなる組合せに対して，最小支持度を $minsup$ として既存の頻出部分グラフマイニングアルゴリズムを適用することにより，その共通一般連結部分グラフ，すなわち頻出部分グラフを求める．

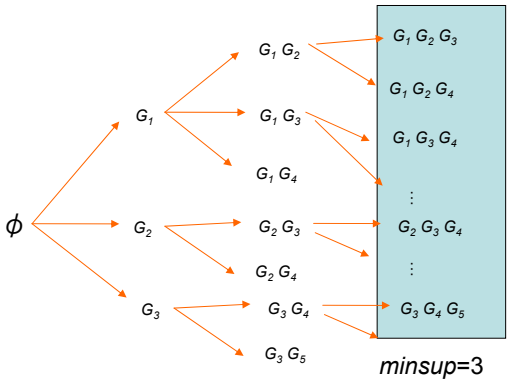

図 2: データグラフ集合 D 内の組合せの列挙木

Input
　D: a set of data graphs whose graph sizes
　　are at least $minsize$
　$minsup$: minimum support
　$minsize$: minimum size of subgraph to find
Step 1. Sort D in the descending order of size;
Step 2. Calculate graph spectrum of each data graph;
Step 3. $FS = \emptyset$;
Step 4. P$(\emptyset, D, minsup, minsize, FS)$;

図 3: 大規模グラフマイニング手法のアルゴリズム

4.2　提案アルゴリズム

提案する大規模グラフマイニング手法，およびその内部で再帰的に用いられる頻出一般連結部分グラフ列挙手続き P のアルゴリズムを図 3，および図 4 に示す．提案する大規模グラフマイニング手法は，図 4 の手続き P を再帰的に呼び出すことにより，データグラフの組合せ D_s を順次列挙し，$|D_s| = minsup$ である D_s に対して，D_s 中のデータグラフに共通する一般連結部分グラフを既存のグラフマイニングアルゴリズムを用いて求める．$|D_s| \leq minsup$ の場合には，補題 3 を用いて D_s が大きさ $minsize$ の共通一般連結部分グラフを持つかどうかを確認する．D_s が大きさ $minsize$ の共通一般連結部分グラフを持たない場合，D_s を包含するデータグラフの組合せは大きさ $minsize$ の共通一般連結部分グラフを持たないと判定できるので，D_s を枝刈りすることができる．

図 4 の手続き P の Step 5 では，新たに発見した頻出一般連結部分グラフ g が以前に発見した頻出部分グラフの集合に含まれるか否かを判定しなければならない．これはグラフ同型判定問題に相当し，その解法には高い計算コストがかかるため，個別にこの問題を解くことは現実的ではない．その為，本研究では手続き P の Step3 で用いるグラフマイニングアルゴリズムとして gSpan を利用し，gSpan により求められる最小 DFS コードを用いて，この判定を行う．

gSpan はデータグラフ中の 1 つの頂点を基準に，1 つの辺を再帰的に深さ優先探索の順序で付加することで，小さな候補頻出部分グラフからより大きな候補頻出部分グラフを順次に列挙し，高速に頻出一般連結部分グラフを発見するアルゴリズムである．同型な頻出部分グラフが重複して発見されることを避けるために，gSpan は各候補グラフに対して，求められる DFS コードの最小性を判定する．DFS コードは辺の追加順序により定まるコードであり，同型な候補グラフでも異なる DFS コードを持ち得る．しかし，最小な DFS コードは一意に決まることから，gSpan ではその DFS コードが最小である候補部分グラフのみを用いることで，重複なく頻出部分グラフを発見する．

Procedure: P$(D_s, D', minsup, minsize, FS)$;
　D_s: a set of data graphs used as a support set
　G: the graph that finally added to D_s
　D': a set of sorted data graphs used for further iterations
　$minsup$: minimum support
　$minsize$: minimum size of frequent subgraphs to find
　FS: a set of frequent subgraphs found so far

Step 1. IF D_s does not have any common subgraph of size
　　$minsize$ based on Interlace theorem
　　then return;
Step 2. IF $|D_s| < minsup$
　　then For each $G_i \in D'$
　　P$(D_s \cup \{G_i\}, D'|i, minsup, minsize, FS)$;
　　// ∗ $D'|i$: A subset of D' consisting of data graphs
　　// whose index is larger than i
Step 3. Find frequent subgraphs S having at least $minsize$
　vertices from D_s using an existing graph mining algorithm;
Step 4. IF $S == \emptyset$ then return;
Step 5. For each $g \in S$
　　IF $g \notin FS$ then $FS = FS \cup \{g\}$
　　ELSE $g.sup = g.sup \cup D_s$;
　　// ∗ $g.sup$: A set of data graphs containing g

図 4: 頻出一般連結部分グラフ列挙手続き P

手続き P では，上記のように gSpan により求められた各頻出部分グラフの最小 DFS コードをその頻度とともに保存する．D が同一であれば，手続き P の Step 3 で gSpan が出力する頻出部分グラフの最小 DFS コードは一意に決まるため，新たな頻出部分グラフ g が発見された場合は，その最小 DFS コードと FS 中の頻出部分グラフの最小 DFS コードを直接比較することで，g が既出か否かを容易に判定ことができる．

5.　期待される効果とまとめ

提案手法はサイズ $minsize$ の共通一般連結部分グラフを持つ可能性がある $minsup$ 個のデータグラフの組合せのみを列挙し，さらにそれらから最小支持度 $minsup$ で頻出部分グラフを求める．このことから，提案アルゴリズムは，小さい候補頻出部分グラフを優先的に探索する従来のアルゴリズムと比較して，より高速にサイズ $minsize$ 以上の大規模な頻出部分グラフを完全検索することが期待できる．

本研究ではグラフスペクトルを用いて大規模頻出部分グラフを完全探索マイニングする大規模グラフマイニング手法を提案した．今後の課題として，様々なデータに提案手法を適用し，実験的に評価することが挙げられる．

参考文献

[Haemers 95] W.H. Haemers, "Interlacing Eigenvalues and Graphs", *Linear Algebra Appl. 226, pp. 593-616, 1995.*

[Ikebe 87] Y. Ikebe, T. Inagaki and S. Miyamoto, "The monotonicity theorem, Cauchy's interlace theorem, and the Courant-Fischer theorem", *American Mathematical Monthly, Vol. 94, Issue 4, pp. 352-354, 1987.*

[Inokuchi 00] A. Inokuchi, T. Washio, and H. Motoda, "An apriori-based algorithm for mining frequent substructures from graph data", *In PKDD'00, pp. 13-23, 2000.*

[Nijssen 04] S. Nijssen and J.N. Kok, "The Gaston Tool for Frequent Subgraph Mining", *Proc. Int'l Workshop on Graph-Based Tools, 127(1) pp. 77-87, 2004.*

[Kuramochi 01] M. Kuramochi and G. Karypis, "Frequent subgraph discovery", *In ICDM'01, pp. 313-320, 2001.*

[Inokuchi 02] Akihiro Inokuchi, Takashi Washio, Kunio Nishimura, and Hiroshi Motoda, "A Fast Algorithm for Mining Frequent Connected Subgraphs", *IBM Research. In IBM Research Report, 2002.*

[Yan 02] X. Yan and J. Han, "gSpan: Graph-Based Substructure Pattern Mining", *In Proc. IEEE Int 'l Conf. on Data Mining ICDM, pp. 721-723, 2002*