様式 C-19

科学研究費補助金研究成果報告書

平成 21 年 4 月 25 日現在

研究種目：基盤研究（C）
研究期間：2007～2008
課題番号：19500089
研究課題名（和文）　数値的・組み合わせ的方法による大規模構造データ検索技術の研究開発

研究課題名（英文）　Research on Searching Large-scale Structural Data with Numerical and Combinatorial Methods

研究代表者
片山　薫（KATAYAMA KAORU）
首都大学東京・システムデザイン研究科・准教授
研究者番号：00336520

研究成果の概要：大規模グラフ集合を効率的に検索するため，数値的方法と組み合わせ的方法による部分グラフ同型性判定(あるグラフが別のグラフに含まれるかどうか判定すること)手法を開発した．行列の固有値に関する Interlace 定理に基づいたグラフ索引手法とグラフフィルタリング手法を提案すると共に，Messmer らの提案したグラフ分解に基づく部分グラフ同型探索アルゴリズムの改良を行った．これらの提案手法について人工的に生成したデータや実際のデータを利用して評価実験を行い，その有効性を検証した．

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	1,800,000	540,000	2,340,000
2008 年度	1,600,000	480,000	2,080,000
年度			
年度			
年度			
総　計	3,400,000	1,020,000	4,420,000

研究分野：データ工学
科研費の分科・細目：情報学・メディア情報学・データベース
キーワード：グラフ，固有値，グラフ分解，構造データ

１．研究開始当初の背景

近年データベースは従来よりも大規模で複雑な構造を持つ情報の管理が求められるようになった．XML や化合物，遺伝子などのデータを計算機で処理する際は，抽象化されラベル付きグラフとして扱われることが多い．そのため，大量のグラフから必要なグラフを効率的に検索するための索引方法(Yan ら(2004))，頻出部分グラフを発見する方法（グラフマイニング）(Huan ら(2004))等様々な研究が進められている．

２．研究の目的

(1)数値的方法による構造データ検索

グラフを検索する際の基本的問題の一つは，あるグラフが別のグラフを内部に含むかどうかを判定すること(部分グラフ同型判定問題)であり，NP 完全であることが知られている．この問題に対して VF2(Cordella ら(2004))や Ullmann(1976)などの組み合わせ的方法によるアルゴリズムが開発されている．ところで，グラフを隣接行列や接続行列として表現すると数値的方法を利用することができる．これにより上記のような組み合

わせ的方法では処理の困難な大きなサイズのグラフについて処理が可能となることが期待される．形の似たグラフを検索することを目的として，これまでにも数値的な方法を利用したグラフの索引手法の研究はあるが(Shokoufandeh ら(1999))，本研究の基礎をなす Interlace 定理(ある行列とその部分行列の固有値に関する事実．以下参照．)に基づくものや，部分グラフ同型判定を直接扱ったものは見られない．

(Interlace 定理)対称行列 A の n 個の固有値を{a(1),a(2),…,a(n)}(a(1)≦a(2)≦…≦a(n))とする．S を $S^tS = I$ を満たす n×m(m<n)の実行列であるとし，対称行列 B を $B = S^tAS$ と定義する．B の m 個の固有値を{b(1),b(2),…,b(m)}(b(1)≦b(2)≦…≦b(m))とすると，a(i)≦b(i)≦a(i+(n-m))(i=1, …,m)が成り立つ．

(2)組み合わせ的方法による構造データ検索
VF2(Cordella ら(2004))や Ullmann(1976)などのアルゴリズムは 2 つのグラフの間の包含関係を判定するものだが，一方で Messmer ら(2000)はグラフ分解を用いて大量のグラフを同時に検索するための興味深い組み合わせ的アルゴリズムを提案した．しかし我々のこれまでの研究から，この方法が多くのメモリを必要とすることなど必ずしも期待通りの性能を持たないことが分った．我々はこの問題を改善しさらなる性能向上を目指す．Yan ら(2004)は大量のグラフを同時に検索するための索引手法を提案しているが，索引構築のためにグラフ集合から頻出部分グラフを発見する必要があり計算コストが大変高い．

３．研究の方法
(1)数値的方法による構造データ検索
グラフが別のグラフに含まれるかどうかという問題は，行列の視点から見ると，ある行列が別の行列を部分行列として含むかどうかという問題になる．我々はこれまでに行列の固有値に関する以下の定理を利用したグラフの検索方法や索引方法を提案している．グラフを隣接行列で表すと，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ないものを見つけることができる．誘導部分グラフではない一般の部分グラフは，主部分行列とはならないため，このままでは Interlace 定理を利用できない．しかし，Haemers (1995)は，グラフを隣接行列ではなく，接続行列を利用した行列として表現すると，部分グラフを主部分行列とすることができることを指摘している．我々はこれらの事実を部分グラフ同型判定処理の前処理として利用する．Interlace 定理を使って検索対象のグラフをフィルタリングすることで部分グラフ同型判定処理を効率化できると共に，固有値の計算に関する様々な知見を利用することができる．固有値計算における知見を利用した処理の効率化や索引方法の改良など様々な課題が残されており，本研究ではそれらの解決を目指す．

(2)組み合わせ的方法による構造データ検索
Messmer らは，あるグラフに含まれるものをグラフ集合の中から効率的に検索する方法を提案した．そのアイデアは，検索対象のグラフ集合について，事前にそれぞれお互いの共通部分を，計算コストが大きくなり過ぎない範囲でできるだけ発見しておくことである．これによって共通部分に対する処理の繰り返しを避けることができる．これまでの研究から Messmer らの方法では大量のメモリが必要となり，あまり大規模なデータを処理できないことが分った．その原因の一つはグラフ分解の際に連結性を考慮しないことにあった．我々はこの点を改良すると共に，グラフ分解によって得られた構造を利用した処理の効率化方法を提案し，実験によりその有効性を確認している．本研究ではさらに，部分グラフ同型判定処理方法の改良などによりこの手法の改良を行う．

４．研究成果
(1)数値的方法による構造データ検索
①二分法を利用した固有値比較手法
グラフの固有値が interlace することを調べる場合，グラフの固有値の大小関係を比較することになる．その際，個々の固有値について厳密な値を計算する必要はなく，大小関係が比較できる程度まで固有値の含まれている範囲が求められればよい．そこで，二分法を利用して二つのグラフの固有値を並行して求めながら，固有値の比較のために必要な精度まで固有値を求めるアルゴリズムを開発した．この手法では，固有値が interlace しないと分かった時点で処理を終えることができるので固有値計算のコストを減らすことができる．実験により，全固有値を計算した後に Interlace 定理を用いる手法より 2 分法に基づく Interlace 定理の利用手法の方が処理が速いことを確認した．

②ラベル付きグラフのフィルタリングのための行列サイズ縮小手法
グラフの固有値を計算する際に必要な行列のサイズを縮小する事で，計算コストを減らすと共に，Interlace 定理による部分グラフ同型性の判定精度を改善する事ができる．部分グラフ に使用されているラベルは，元のグラフに全て含まれている事を利用して，部

分グラフではないグラフをフィルタリングする．また，部分グラフに含まれないラベルを持つ頂点と枝は，部分グラフ同型性判定に影響を与えないため削除することにより，行列サイズを縮小することができる．人工的に生成したデータを用いた評価実験により，提案手法が有効である事を確認した．

③固有値を利用した構造データの索引手法 1
Interlace 定理を用いることで，あるグラフが別のグラフの誘導部分グラフではないことを数値比較によって判定することができる．しかし，この方法では部分グラフでないことしか判定できないため，改めて組み合わせ的な方法で部分グラフ同型判定を行う必要がある．我々は，Interlace 定理と固有値に基づく 2 分木を用いた索引と，多分木を用いた索引を提案した．2 分木を用いた索引は各ノードにおいてグラフの固有値の中央値を求め，その値に従ってグラフを 2 分割する．これを木の高さだけ繰り返し，分割されたグラフの部分集合を葉に格納する．グラフ問い合わせの際，各ノードに格納されている中央値と対応する問い合わせの固有値とを比較しながら木を探索し，辿り着いた葉に含まれるグラフを候補グラフとする．これにより，問い合わせを含まないグラフを少ない判定回数で除去することが可能である．多分木を用いた索引はデータ構造を 3 分木以上に拡張した手法であり，同じ条件であれば 2 分木を用いたインデクスよりも多くのグラフを除去することが可能である．

④固有値を利用した構造データの索引手法 2
2 分木や多分木を用いた索引手法では Interlace 定理の条件式の一部しか利用していないため，索引の探索のみではグラフをほとんど除去できない場合があった．そこで，Interlace 定理の条件式の逐次的判定を必要としない新たな索引手法を開発した．この索引では Interlace 定理によるグラフの包含関係に基づいてグラフが格納されたノードを連結する．あるグラフの固有値が別のサイズの小さいグラフの固有値と Interlace 定理の条件式を満たすとき，そのグラフは別のグラフに interlace されるという．データベース中のグラフｇとｈについて，ｇがｈに interlace されるならｈをｇの子ノードとする．この処理をデータベース中のすべてのグラフについて行うことでデータ構造を構築する．問い合わせを用いてデータ構造を探索することで，逐次的に Interlace 定理の条件式を判定するよりも効率的に処理することが可能である．提案した手法を評価するために，実データである AIDS Antiviral Screen データと，人工的に生成したデータを用いて実験を行った．AIDS Antiviral Screen データについては，Cheng ら(2007)によって提案されたグラフ索引手法である FG-Index と比較した．実験の結果，提案手法は FG-Index に処理時間の面では劣っているが，FG-Index が扱うことができないグラフを処理できることを確認した．提案手法が 2 分木や多分木を用いたインデクスよりも有効であることも確認した．また，提案手法を利用することで，部分グラフ同型判定の組み合わせ的手法である VF2 だけを用いる場合よりも効率的に部分グラフを判定することが分かった．

(2)組み合わせ的方法による構造データ検索
①グラフ分解による非連結グラフの効率的処理
入力グラフ(問い合わせグラフ)とモデルグラフ集合(グラフデータベース)内の各グラフを連結グラフに限定し，さらにグラフ分解時に常に連結グラフに分解すると，検索時のマッチングパターンの増加を抑えることができる．これによって，より大規模なグラフ集合の処理が可能となるが，入力グラフ，モデルグラフが連結グラフに限定されてしまうため，Messmer らの手法では可能であった非連結グラフの部分グラフ同型判定ができない．そこで，入力グラフ，モデルグラフが非連結である場合に対応できるようモデルグラフが非連結グラフの場合，入力グラフが非連結グラフの場合それぞれについて処理を追加した．モデルグラフが非連結グラフの場合，モデルグラフの各連結成分の入力グラフに対する部分グラフ同型判定については既存の手法を利用できる．しかし，あるモデルグラフの連結成分が全て入力グラフに含まれていても，そのモデルグラフが入力グラフに含まれているとは限らないため，各連結成分の部分グラフ同型判定結果を使って元のモデルグラフが入力グラフの部分グラフであるかどうかを判定する処理を追加する．入力グラフが非連結グラフの場合，入力グラフに枝を追加して連結グラフにする．連結グラフになった入力グラフとモデルグラフ集合とで部分グラフ同型判定を行い，入力グラフの部分グラフと判定されたモデルグラフについては追加した枝を含むかどうかを調べることで正しい判定結果を得る．このとき，枝のラベルの付け方を工夫することで処理を高速化する．モデルグラフ集合で用いられる枝のラベルが既知である場合には，入力グラフに追加する枝にモデルグラフ集合内で使用されていないラベルをつける．これにより，入力グラフの部分グラフであると判定されたグラフが追加された枝を含むことは起こり得ず，そのグラフは確実に枝を追加する前の入力グラフにも含まれることになる．人工データを用いた評価実験により，提案手法

の有効性を検証した.

②再帰的な処理による部分グラフ同型性判定処理の効率化
Messmer らの手法では，モデルグラフ集合内の各グラフを再帰的に分解してできるデータ構造を予め構築するが，その際各グラフ間の共通部分がデータ構造内で共有される．このデータ構造を利用して誘導部分グラフ同型問い合わせを処理する．Messmer らのアルゴリズムでは，下位のグラフと問い合わせグラフとの誘導部分グラフ同型から，上位のグラフと問い合わせグラフとの誘導部分グラフ同型を発見し，データ構造内の全てのグラフをボトムアップに処理する．我々は，この処理を再帰化することによって冗長な計算を削減する手法を開発した．すなわち，あるグラフから問い合わせグラフへの誘導部分グラフ同型を返すアルゴリズムは，その子に対する同じアルゴリズムによる問い合わせの結果を用いて答えを返す．このとき，子の一方が誘導部分グラフ同型を一つも返さなければ，もう一方に対する問い合わせは行わない．この再帰的な処理により，必要のないノードへの問い合わせが無くなり，冗長を削減することができる．モデルグラフ集合内に問い合わせグラフへの誘導部分グラフ同型を一つも持たないグラフが多いほど，多くの冗長ノードが多いため，削減効果は高くなる．処理の再帰化によって，さらに局所的な問い合わせ（モデルグラフ集合の任意の部分集合内のグラフから問い合わせグラフへの誘導部分グラフ同型を検出する問題）にも効率的に対応することが可能になる．Messmer らの手法では，たとえ局所的な解しか必要としていなくても，通常の問い合わせと同じように，データ構造内の全てのノードを処理しなければ，局所的な問い合わせに対応できない．しかし，再帰的なアルゴリズムであれば，必要なノードのみ問い合わせることができ，より効率的な処理が可能になる．評価実験において，化合物データベースおよび人工データを用いて提案手法と Messmer らの手法，一対一の誘導部分グラフ同型検出による逐次処理を比較し，提案手法の有効性を示した．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計 11 件）
① 大河原達郎，片山薫，Hyperlinkedness に基づく Hypertree Decomposition 構築アルゴリズム，第 1 回データ工学と情報マネジメントに関するフォーラム (DEIM2009)，2009 年 3 月 10 日，静岡県掛川市.
② 関建二郎，片山薫，再帰的な処理によるグラフ分割を用いた部分グラフ問い合わせ処理の効率化，第 1 回データ工学と情報マネジメントに関するフォーラム (DEIM2009)，2009 年 3 月 10 日，静岡県掛川市.
③ 関建二郎，片山薫，グラフ分解を用いた構造による部分グラフ問い合わせ処理の改良とその実験的評価,iDB フォーラム (iDB2008)，2008 年 9 月 22 日，福島県福島市.
④ 森垣潤一，片山薫，高次特異値分解の画像分類への応用，電子情報通信学会 第 19 回データ工学ワークショップ (DEWS2008)，2008 年 3 月 11 日，宮崎県宮崎市.
⑤ 大河原達郎，片山薫，Hypertree Decomposition の正規形を構築するためのアルゴリズム，電子情報通信学会 第 19 回データ工学ワークショップ (DEWS2008)，2008 年 3 月 11 日，宮崎県宮崎市.
⑥ 布施貴義，片山薫，非連結グラフに対応した大量グラフ集合からのグラフ分解による部分グラフ発見手法，電子情報通信学会 第 19 回データ工学ワークショップ (DEWS2008)，2008 年 3 月 10 日，宮崎県宮崎市.
⑦ 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，電子情報通信学会 第 19 回データ工学ワークショップ(DEWS2008)，2008 年 3 月 10 日，宮崎県宮崎市.
⑧ 高橋俊介，片山薫，グラフ索引のための Interlace 定理によるグラフの包含関係を考慮したデータ構造の提案，電子情報通信学会 第 19 回データ工学ワークショップ(DEWS2008)，2008 年 3 月 10 日，宮崎県宮崎市.
⑨ 高橋俊介，片山薫，Interlace 定理に基づく多分木を用いたグラフの索引手法，夏のデータベースワークショップ (DBWS2007)，2007 年 7 月 4 日，宮城県仙台市.
⑩ 長屋未来，片山薫，ラベル付きグラフのフィルタリングのための行列サイズ縮小手法，夏のデータベースワークショップ (DBWS2007)，2007 年 7 月 4 日，宮城県仙台市.
⑪ Takayoshi Fuse, Shotaro Nishimura, Kaoru Kayayama, Enabling the Messmer's Graph Decomposition Approach for Subgraph Isomorphism Detection to Apply to a Larger Set of Graphs, The 2007 International Conference on Data Mining (DMIN2007), 2007 年 6 月 28 日, Las Vegas, Nevada, USA.

〔その他〕
電子情報通信学会 第 19 回データ工学ワークシップ最優秀論文賞受賞(学会発表⑧)

６．研究組織
(1)研究代表者
片山　薫(KATAYAMA KAORU)
首都大学東京・システムデザイン研究科・准教授
研究者番号：00336520
(2)研究分担者

(3)連携研究者