様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年５月２２日現在

機関番号：２２６０４
研究種目：基盤研究（C）
研究期間：2009～2011
課題番号：２１５００１０４
研究課題名（和文）　大規模構造データのための数値的手法を利用した検索技術の研究開発

研究課題名（英文）　Study on Efficient Retrieval Methods for Large-scale Structured Data by Numerical Approaches

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

研究成果の概要（和文）：グラフやその拡張である階層グラフ、テンソルデータなど様々な構造を持つ大規模なデータを数値的に扱うことにより、利用者の求める部分構造を含むデータを、組合せ的な方法だけを利用する方法よりも効率的に検索する技術の研究開発を行った。人工データや実際のデータを用いた評価実験により、構造データを組合せ的な方法を用いたアプローチよりも大規模なデータを効率的に処理することができる場合を示した。

研究成果の概要（英文）：We develop efficient methods for retrieving substructures of large-scale graphs, hierarchical graphs and tensor data by numerical approaches. We evaluate the proposed methods experimentally with real data and artificial data, and show the cases where they are efficient in comparison with searching methods using only combinatorial approaches.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００９年度	1,200,000	360,000	1,560,000
２０１０年度	1,100,000	330,000	1,430,000
２０１１年度	1,100,000	330,000	1,430,000
年度			
年度			
総　計	3,400,000	1,020,000	4,420,000

研究分野：総合領域
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：部分グラフ、固有値、interlace、HOSVD、PARAFAC、hypertree decomposition

１．研究開始当初の背景

近年インターネット上でのデータ交換形式として普及の進む XML(Extensible Markup Language)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(Cheng ら(SIGMOD2007))や、利用者のとって重要なグラフパターンを発見する方法(Yan ら(SIGMOD2008))等データ工学の分野においても様々な研究が進められている。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)であるが、NP 完全であることが知られており、組み合わせ的方法で大規模なグラフを扱うことは容易ではない。そこで我々は、行列の固有値に関する定理(Interlace 定理)を利用したグラフの検索方法や索引方法の研究を進めており、電子情報通信学会第 19 回データ工学ワークショップでは最優秀論文賞(高橋ら

(DEWS2008))を受けるなど成果を得ているが、研究計画に述べるような様々な課題が残されている。本研究ではそれらの解決を目指すとともに、テンソルデータなど他の構造データへの応用手法について研究を行う。類似した(形の似た)グラフの検索を目的として、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(Shokoufandeh ら(CVPR1999))はあるが、Interlace 定理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他に見られない。

Messmer ら(TKDE2000)は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。我々はこの方法が多くの記憶領域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を確認し、使用する記憶領域量を大幅に削減する方法などの開発を行った。我々の開発した処理手法は現在最も効率的なアルゴリズムの一つである VF2(Cordella ら(2004))との比較においても優れていることを様々な実験において検証している(布施ら(DEWS2008), Fulse et al.(DMIN2007))。しかし処理対象となるデータは日々増大しており、より効率的な処理手法の開発が求められている。開発中のシステムのアルゴリズムを改良し、処理の効率化を目指す。

代表的な文書処理手法の一つである潜在的意味解析(Latent Semantic Analysis)では、文書集合中の文書と文書に含まれる単語の対応を行列として表現したものを特異値分解することによって文書間の関連などを求めている。De Lathauwer ら(2000)によって提案された高次特異値分解(Higher-Order Singular Value Decomposition(HOSVD))は、特異値分解をより高次の「行列」(テンソルと呼ばれる)向けに拡張したものであり、顔画像認識や手書き文字認識、グラフデータ処理(Sun ら(KDD2006))などに応用され近年研究が進められている。テンソルもまた構造化されたデータの一種であり、我々は HOSVD を利用した画像の分類システムを構築し実験的に性能を評価した(森垣ら(DEWS2008))。今後さらに画像の特徴量の選択方法や分類アルゴリズムの改良などにより、性能の向上が可能であると考えられる。

２．研究の目的

本研究では、大規模なグラフやテンソルデータ(多次元データ)などの構造データを効率的な処理を可能にするため、数値的方法と組み合わせ的方法の双方の利点を活かした手法を開発する。これにより、組み合わせ的な方法だけでは処理が困難な規模を持つ構造データの処理することや、組み合わせ的な方法のみを利用した場合よりも処理を効率化することができるようになるものと期待される。我々は平成 19 年度～平成 20 年度において受けた科学研究費補助金基盤研究(C)(一般)「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」により本アプローチによる研究を進めており、本研究ではこれまでの研究から明らかになった課題の解決するとともに、新たな手法の開発によって研究開発をさらに進展させることを目指す。

３．研究の方法

(1) 固有値を利用した構造データ検索
グラフは隣接行列、接続行列などの行列として表現することができる。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。対称行列とその部分行列の固有値には Interlace 定理(Haemers(1995))として知られる関係がある。本研究では、Interlace 定理や行列の固有値を求める数値計算手法を応用し大規模な構造データの効率的検索手法を開発する。

(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。Messmer らの提案は処理中に大量の記憶領域が必要になるという問題があることが分かったため、我々はその点を改善し処理を効率化したアルゴリズムを提案した(Fuse et al.(DMIN2007))。本研究では、提案手法のさらなる改良を行う。

(3) テンソルデータ処理手法の改良
データの類似検索に利用される代表的なテンソル分解方法には、HOSVD と PARAFAC がある。多様な実験データを利用した従来の画像分類手法との比較等によって、これらの長所と短所を明らかにすると共に、その改良を行う。

(4) hypertree decomposition 構築アルゴリズムの改良
hypertree decomposition は Gottlob らによって提案されたハイパーグラフの分解方法である。これを利用することによ

りハイパーグラフ上の問題を効率的に処理することができる。グラフに対する同様の分解方法である tree decomposition を構築するアルゴリズムを応用することにより、hypertree decomposition を効率的に構築するアルゴリズムを開発する。

(5) 位相限定相関法を用いた構造データ検索方法の開発
位相限定相関は主に画像等の二次元データにおける検索方法として用いられている。コンピュータ断層撮影(CT:Computed Tomography)で用いられる技術と組み合わせることにより、一般的な構造データの検索手法への応用を目指す。

４．研究成果

(1) 固有値を利用した構造データ検索
組合せ的な方法では処理が困難な大規模グラフの効率的な検索を実現するため、グラフの固有値を利用して検索対象でないグラフをフィルタリングする手法について研究を行った。大規模な行列の固有値を求めるために開発された様々な数値計算技術を応用することによって、今後さらに大規模な構造データの処理が可能になるものと期待される。

① ラベルによるグラフ分解によるフィルタリング精度と処理能力の向上
この手法は問合せグラフと検索対象のグラフのサイズの差が大きいとフィルタリング精度が低くなるという問題があった。そこで、グラフをラベルによって分解することによりこの問題の影響を低く抑えると共に、グラフの固有値をより効率的に計算する手法を開発した。また、グラフのラベルによる分解を利用してこれまで開発を進めてきたグラフの索引手法を改良した。

② サブグラフ問合せとスーパーグラフ問合せに対応したグラフ索引の開発
グラフデータベースに対する典型的な問合せには、データベース中のグラフから問合せグラフのサブグラフの発見を求めるものと、スーパーグラフの発見を求めるものがある。これまでに提案された索引の多くは、どちらか一方の問合せだけを処理できるものであった。それに対し我々は、グラフの固有値を利用して、一つでどちらの問合せも処理することができる索引構造を開発した。

③ 固有値を利用したグラフ索引方法の改良
固有値の間の関係をより詳細に調べて索引を構築することにより、従来提案した索引よりも効率的なグラフ検索を実現する方法を開発した。

(2) 固有値を利用した階層構造を持つグラフの検索手法の開発
グラフは様々な実際問題のモデルとして用いられているが、蛋白質の構造やソーシャルネットワークなどのより複雑な構造を表現するには不十分である。このような複雑な構造を表現するため、グラフの各頂点が別のグラフを含む階層的な構造を持つグラフ(階層グラフ)と、その部分構造である部分階層グラフを定義した。その上で、グラフの固有値を用いて問合せを部分階層グラフとして含む、データベース中の階層グラフを効率的に検索する手法を開発した。人工的に生成されたデータだけではなく、テキスト検索で用いられるデータ(拡張固有表現階層)を利用してその効率について評価を行った。このような複雑な構造を持ったデータに対する効率的な検索技術が今後さらに求められるようになるものと思われる。

(3) テンソルデータに対する分解手法の改良
テンソルデータはデータを行列で表現する方法の自然な拡張であり、今後さらに様々な用途で利用されるものと期待される。

① 相互部分空間法に基づく HOSDV によるテンソルデータ分類性能の向上
Savas らは高階特異値分解(HOSVD)(Lathauwer ら)で得られた結果を基に「直交基底行列」を求め，未知の入力データとのスカラー積を類似度とする精度の高い手書き数字分類手法を提案した。我々はこの手法を Kohonen らによって提案された学習部分空間法及び、Sun らによって提案された組み合せ型部分空間構成法を用いて拡張し、テンソルデータを分類する手法を開発した。評価実験により Savas らの手法と比較してパラメーターの設定により精度が向上する場合があることを確認した。

② 相互部分空間法に基づく PARAFAC によるテンソルデータ分類方法の改良
テンソルデータの代表的な処理手法の一つである PARAFAC は、よく知られたパターン認識手法である部分空間法の拡張とみなすことができる。このことを利用し、部分空間法の拡張である相互部分空間法のアイデアを用いて PARAFAC を改良し、画像分類に応用してその性能を評価した。

(4) 効率的な Hypertree Decomposition 構築アルゴリズム
Gottlob らによって提案されたハイパーグラフの分解手法である Hypertree Decomposition を効率的に構築するアルゴリズムを開発した。関係データベースにおける Query Containment 問題や、人

工知能における制約充足問題はハイパーグラフとして表現することができ、一般にはNP完全であることが知られている。本研究はこれらの問題を効率的に処理するために利用することができる。

(5) 位相限定相関法を用いた構造データ検索手法に関する研究
コンピュータ断層撮影(CT:Computed Tomography)では構造データ(体の断面)に関する情報は間接的な形(投影データ)で与えられている。本研究では、このようなデータベース内のデータと問合せデータが共に投影データの形で与えられる場合を対象として、位相限定相関法を用い問合せを部分構造として含むデータを効率的に検索する手法について研究を行った。今後CTから得られたデータやセンサーから得られたデータ等対象を直接表現していなデータはさらに増加し、それらに対する効率的な検索技術が求められるものと考えられる。

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９件）

① 鈴木純，片山薫，階層グラフ発見手法の効率化とテキスト検索への応用，第4回データ工学と情報マネジメントに関するフォーラム（DEIM2012），2012年3月4日，兵庫県神戸市.
② 三井良太，片山薫，固有値を利用したインデクス手法の改良，第4回データ工学と情報マネジメントに関するフォーラム(DEIM2012)，2012年3月3日，兵庫県神戸市.
③ 宮奥祥多，三井良太，片山薫，サブグラフ問い合わせとスーパーグラフ問い合わせに対応した固有値を用いたグラフインデクス手法，第3回データ工学と情報マネジメントに関するフォーラム(DEIM2011)，2011年2月27日，静岡県伊豆市.
④ 鈴木純，片山薫，階層構造を持つグラフの固有値を利用した発見手法，第3回データ工学と情報マネジメントに関するフォーラム（DEIM2011)，2011年2月28日，静岡県伊豆市.
⑤ 根本祐介，片山薫，位相限定相関法を用いた大規模高次元データ検索手法の提案，第3回データ工学と情報マネジメントに関するフォーラム（DEIM2011)，2011年2月28日，静岡県伊豆市.
⑥ Kaoru Katayama, Tatsuro Okawara, Yuka Ito, A Greedy Algorithm for Constructing a Low-Width Generalized Hypertree Decomposition, 13th International Conference on Database Theory (ICDT2010), 2010年3月24日，ローザンヌ(スイス).
⑦ 森垣 潤一，片山 薫，HOSVDにおける自己相関行列の補正を用いた高階データ分類手法，第2回データ工学と情報マネジメントに関するフォーラム（DEIM2010)，2010年2月28日，兵庫県淡路市.
⑧ 宮奥 祥多，片山 薫，ラベルによるグラフ分割を用いた固有値に基づくグラフ索引手法の改良，第2回データ工学と情報マネジメントに関するフォーラム(DEIM2010)，2010年2月28日，兵庫県淡路市.
⑨ 天笠 暢甫，片山 薫，ラベルと頂点の次数を用いたグラフフィルタリング手法の改良，第2回データ工学と情報マネジメントに関するフォーラム（DEIM2010)，2010年3月1日，兵庫県淡路市.

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
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：