様式 C-19

科学研究費補助金研究成果報告書

平成　23 年　5 月　24 日現在

機関番号：32601
研究種目：若手研究(B)
研究期間：2009～2010
課題番号：21700167
研究課題名（和文）　グラフ閉包演算を用いた頻出飽和部分グラフマイニングの実現とその並列化
研究課題名（英文）Closed Frequent Subgraph Mining by Graph Closure Operation and Its Parallelization
研究代表者
大原　剛三（OHARA KOUZOU）
青山学院大学・理工学部・准教授
研究者番号：30294127

研究成果の概要（和文）：本研究では，与えられた部分グラフを包含する飽和部分グラフ（その部分グラフを含むグラフのうち，同じ頻度をもち，かつ最大であるグラフ）を計算するグラフ閉包演算，および同型な部分グラフを効率的に判定する手法を提案し，グラフデータベース中に一定頻度以上の割合で出現する頻出飽和部分グラフを効率よく列挙する頻出飽和部分グラフマイニングシステムを実現するとともに，その並列化プロトタイプシステムを構築した．

研究成果の概要（英文）：In this work, we proposed the graph closure operator, as well as an efficient method for graph isomorphism checking that takes advantage of occurrences of a graph pattern. Given a subgraph, the operator generates a closed subgraph including it, where the closed subgraph is the maximum subgraph among those which have the same frequency as the given one. Then, we developed an efficient method for enumerating all possible closed frequent subgraphs in a graph database by means of them, and parallelized its inner processes to improve its scalability.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	1,500,000	450,000	1,950,000
2010 年度	1,800,000	540,000	2,340,000
年度			
年度			
年度			
総　計	3,300,000	990,000	4,290,000

研究分野：知識発見，知識工学
科研費の分科・細目：情報学・知能情報学
キーワード：グラフマイニング，データマイニング，機械学習，パターン発見

１．研究開始当初の背景

計算機と情報通信ネットワークの急速な普及に伴い，様々な分野において膨大な電子化情報が蓄積され，対象データも分子構造などの複雑な構造を有するものが多用されつつある今日，それら構造をもつデータから有用な知識を発掘するデータマイニング技術への要求が高まっている．特に，複雑な構造をもつデータは頂点，及び頂点間を結ぶ辺により構成されるグラフとして表現できる為，近年，グラフ構造から有意な塊（部分グラフ）を発掘するグラフマイニングが盛んに研究されており，国内外において効率的な多数のアルゴリズムが提案されている．

それらの多くは，所与のグラフ集合に一定頻度（最小頻度閾値）以上で現れる頻出部分グラフを効率的に発見するが，発見できる頻出グラフの大きさは最大でも頂点数が数十程度までであり，現実的な時間内で解を得るには，潜在的に大きな頻出部分グラフを含み

得る大規模なグラフへの適用は困難である．これは，大きな頻出部分グラフが存在する場合，その全ての部分グラフもまた頻出部分グラフとなり，その数は頂点数に対して指数オーダーとなる為である．膨大な頻出部分グラフ全てを列挙する事は現実的ではない．この為，原子数が 1,000 個以上となる高分子や，数百以上の遺伝子間の相互関係を表現する遺伝子ネットワークの解析に対するグラフマイニングへの要求が高まっている反面，それらのデータへの既存のグラフマイニング技術の適用は困難となっている．

この問題に対し，頻出部分グラフではなく同一頻度の頻出部分グラフのうち包含関係において極大である頻出飽和部分グラフを列挙することが考えられ，その為のアルゴリズムが提案されている．任意の頻出部分グラフは，それ自身が頻出飽和部分グラフであるか，いずれかの頻出飽和部分グラフの部分グラフとなる為，頻出飽和部分グラフを全て発見できれば，全ての頻出部分グラフを再現できる．しかし既存手法は頻出部分グラフを列挙するアルゴリズムを基礎としている為，問題の根本的な解決とはなっていない．

２．研究の目的

以上のような背景から，本研究では，より大規模なグラフに適用可能であり，かつデータの大規模化にロバストな頻出飽和部分グラフマイニングシステムを実現することを目的とする．その為に，以下の部分目標を設定する．

（1）グラフ閉包演算を実現する効率的なアルゴリズムの実現

頻出飽和部分グラフが，ある部分グラフの閉包（その部分グラフを内包し，かつ頻度がその部分グラフと同一である極大な部分グラフ）である点に着目し，頻出飽和部分グラフのみを効率的に列挙するために，与えられた部分グラフの閉包を計算するグラフ閉包演算を定義し，それを効率的に実現するアルゴリズムを開発する．

（2）効率的なグラフ同型性判定手法の実現

最終的な実現目標である頻出飽和部分グラフ列挙アルゴリズム全体を通して同一の頻出飽和部分グラフを重複して列挙することを回避する為，2 つのグラフが同型であるかどうかを効率的に判定する手法を実現する．グラフの頂点数に対して指数オーダーの計算コストを要する既存手法に対して，グラフの頂点と辺の総数に対して多項式時間で計算可能なグラフマイニングに特化した手法の実現を目指す．

（3）グラフ閉包演算に基づく頻出飽和部分グラフマイニングアルゴリズムの並列化

上記 2 つの目標を達成することで得られた技術を統合して頻出飽和部分グラフマイニングアルゴリズムを実現するとともに，対象とするグラフが大規模化し，頻出飽和部分グラフ数自体が多くなる場合においても，提案アルゴリズムの実行効率を極力落とさない為に，アルゴリズムを並列化し，対象データの大規模化に一定のロバスト性を有するシステムの実現を目指す．

３．研究の方法

前述の各研究目標に対する本研究におけるアプローチについて述べる．

（1）グラフ閉包演算を実現する効率的なアルゴリズムの実現について

ここでは，グラフ閉包演算が与えられた部分グラフ g を含むグラフ集合 S の共通部分グラフの中で g を包含する極大なものを見つける事に等しい点に着目し，そのような共通部分グラフを効率的に発見することで，効率的なグラフ閉包演算を実現する．

（2）効率的なグラフ同型性判定手法の実現について

一般に，グラフマイニングにおいては，頂点の参照順序に依存して同一の部分グラフを複数回列挙する場合が生じ得る．その重複列挙を回避する為には，グラフパターン同士の同型性を判定する必要があり，その為にはグラフの頂点数に対して指数オーダーの計算時間を要する．これに対し，本研究では，部分グラフがパターンとして同型であれば，そのパターンに対応するデータ中に実在するすべての部分グラフ（そのグラフパターンの出現と呼ぶ）も一致することに着目し，事例の代表元を一意に定め，その代表元を比較することでグラフの頂点と辺の総数に対して多項式オーダーで計算可能なグラフ同型性判定を実現する．

（3）グラフ閉包演算に基づく頻出飽和部分グラフマイニングアルゴリズムの並列化について

グラフ閉包演算の内部処理を並列化し，利用 CPU コア数を増やすことによる台数効果により，対象データの大規模化に対してロバストなシステムの実現を目指す．

４．研究成果

（1）グラフ閉包演算を実現する効率的なアルゴリズムの実現に関する成果

与えられた部分グラフ g を含むグラフ集合 S の共通部分グラフの中で g を包含する極大なものを効率的に見つけるために，新たなグラフの表現方法と，その表現方法に基づく g を含む極大な S の共通部分グラフ，すなわち g の閉包を効率的に列挙するグラフ閉包演算を実現した．新たなグラフ表現では，与えられたデータ中で部分グラフ g に実際に接続し得る辺と，各辺が S 中のどのグラフに含まれるかの情報をグラフパターンとと

もに保持する．この情報は，g の頻度計算の際には必ずデータ中の g の出現を列挙する必要があることから，その際に容易に取得可能である．それらの辺を g のオープンエッジと呼び，以下，$OE(g)$ と表記する．

一方，提案したグラフ閉包演算はこの $OE(g)$ を用いて g の閉包を求める．そのアルゴリズムの概要を以下に示す．

1. 辺の集合 Ea を空集合に初期化する．
2. g の各オープンエッジ $e \in OE(g)$ に対して以下を実行．
   2.1. g に辺 e を加えたものを g' とする．
   2.2. $e' \neq e$ であるすべての $e' \in OE(g)$ に関して，$G(e) \subseteq G(e')$ であれば，Ea に e' を加える．
   2.3. Ea が空集合でなければ，g' に Ea 中の辺を加えたものを新たな g として 2.2 を実行．そうでなければ，2.4 へ．
   2.4. g' を g の閉包として出力．

上記において，$G(e)$ はオープンエッジ e を実際に含むグラフの集合を表すものとする．言い換えると，それは元の部分グラフ g とそれに接続する辺 e からなるグラフパターンを含むグラフの集合となる．

グラフデータベース中に一定頻度以上で出現する頻出部分グラフは，頻出飽和部分グラフの部分グラフとなっていることから，すべての頻出飽和部分グラフを効率よく列挙することで，総数がはるかに多いすべての頻出部分グラフを効率よく発見することが可能であり，それを実現する上記アルゴリズムはグラフマイニングの適用範囲をさらに広げることになり，その意味において意義の高いものである．

（2）効率的なグラフ同型性判定手法の実現に関する成果

グラフパターン g が与えられた際に，データ中の g のすべての出現を，グラフ・頂点・辺の ID 番号，頂点ラベル，辺ラベルに基づき辞書式順序に並べ，その先頭となるものを g の出現の代表元とし，代表元同士を比較することでグラフパターンの同型性を判定する手法を実現した．グラフパターンの生成過程において，同型な 2 つのパターン上の頂点と，その出現におけるデータ中の頂点の対応関係が異なっていても，出現自体は一致するため，同型であるグラフパターンは同一の代表元を必ずもつ．部分グラフ列挙の過程では，新たなグラフパターンが見つかる度にその代表元を接頭辞木に登録する．具体的には，各代表元はその頂点 $v_1,\cdots,v_n$，辺 $e_1,\cdots,e_m$ を順に列挙した形で表現され，各頂点，辺を接頭辞木の節点に対応させる．ここで，各頂点，辺はその ID 番号とラベルに基づき辞書式順序に並んでいるものとする．このとき，同型性判定はその接頭辞木を辿ることによって実現でき，その計算コストは代表元の頂点，辺の並び替えと接頭辞木を辿るコストに比例する．

この同型性判定法は，一般のグラフ同型性判定には直接適用することはできないが，グラフパターンを逐次的に列挙する他のグラフマイニングアルゴリズムにおいてはパターンの重複列挙を回避するために利用可能であり，既存アルゴリズムの効率化に寄与し得るものとしても意義の高いものである．

（3）グラフ閉包演算に基づく頻出飽和部分グラフマイニングアルゴリズムの並列化に関する成果

まず，上記で述べたグラフ閉包演算とグラフ同型性判定法を用いた頻出飽和部分グラフマイニングアルゴリズムを実現した．このアルゴリズムは，グラフデータベース中に出現し得る頂点パターンを初期グラフパターンとし，深さ優先探索順序で再帰的にグラフ閉包演算を適用することにより頻出飽和部分グラフを列挙する．その際，先のグラフ閉包演算アルゴリズムのステップ 2.4 にグラフ同型性判定を組込み，同型な頻出飽和部分グラフが複数回列挙されるのを回避する．ベンチマークデータを利用した評価実験では，頻出部分グラフを列挙する従来アプローチよりも効率的に全頻出飽和部分グラフを列挙できることを確認した．

さらに，グラフ閉包演算の内部処理を並列化するというアプローチを取り，利用 CPU コア数を増やすことによる効率化を図った．人工的に生成したグラフデータ，化合物データ，および社会ネットワーク上の情報伝搬過程から取得したネットワーク構造などを用いて行った評価実験では，一定の効率改善を確認した．ただし，CPU コア数に対して比例するほどの台数効果が必ずしも得られない結果となった．これは，内部処理の独立性が必ずしも高くなかったこと，および実装上の技術的な問題に起因したものであり，今後，改善が可能である．並列化による一定の効率改善は確認できており，頻出飽和部分グラフマイニングの大規模データへの適用に向けての道筋をつけるものとしてここで得られた知見は意義のあるものである．

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 3 件）

① Kazumi Saito, Masahiro Kimura, Kouzou Ohara, and Hiroshi Motoda, Selecting Information Diffusion Models over Social Networks for Behavioral Analysis, The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in

Databases 2010 (ECML/PKDD2010), 2010 年 9 月 22 日, スペイン・バルセロナ自治大学.

② Kazumi Saito, Masahiro Kimura, Kouzou Ohara, and Hiroshi Motoda, Behavioral Analyses of Information Diffusion Models by Observed Data of Social Network, The International Conference on Social Computing, Behavioral Modeling, and Prediction (SBP10), 2010 年 3 月 31 日, アメリカ・アメリカ国立衛生研究所.

③ Kazumi Saito, Masahiro Kimura, Kouzou Ohara, and Hiroshi Motoda, Learning Continuous-Time Information Diffusion Model for Social Behavioral Data Analysis, The 2nd Asian Conference on Machine Learning (ACML2010), 2009 年 11 月 3 日, 中国・南京大学.

６．研究組織

(1)研究代表者

大原　剛三 (OHARA　KOUZOU)
青山学院大学・理工学部・准教授
研究者番号：30294127