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２４日現在

研究種目：若手研究(B)
研究期間：2007～2008
課題番号：19700146
研究課題名（和文）　階層的構造データからの特徴的パタン発見に関する研究

研究課題名（英文）　Characteristic Pattern Mining in Hierarchical Structured Databases

研究代表者
尾崎 知伸（OZAKI TOMONOBU）
神戸大学・自然科学系先端融合研究環重点研究部・助教
研究者番号：40365458

研究成果の概要：
本研究では，系列やグラフデータなどの組合せである複合的な構造データを対象に，データ中に頻繁に出現する代表的・特徴的パタンを効率的に発見するアルゴリズムの開発を行った．その結果，(1)木構造データを対象とした制約付きパタン，(2)構造データ集合のグラフを対象とした各種特徴的パタン，(3)グラフ系列を対象とした飽和強相関パタンなど，複合構造データに対する新たなパタンの効率的な発見手法の開発に成功した．

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	1,200,000	0	1,200,000
2008 年度	1,400,000	420,000	1,820,000
年度			
年度			
年度			
総　計	2,600,000	420,000	3,020,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：知識発見とデータマイニング

１．研究開始当初の背景

近年，系列や木，あるいはグラフ構造を対象とした構造データマイニング(Structured Data Mining, SDM)の研究が盛んに行われている．しかし，マイニングの対象が必ずしも木構造やグラフとして自然にモデル化できるとは限らず，異種の構造データの組み合わせとして表現されることも少なくない．

本研究では，これらの複合的な構造データを「複合構造データ」と呼ぶ．複合構造データは，その構成から(1)種々の構造データが各次元を構成する「多次元構造データ」，(2)ある構造の中に別の構造が現れる「階層的構造データ」，(3)構造を構成する要素間の関係を表す「関係構造データ」などに大別できる．

既存の SDM 手法は，木やグラフなど，単一の構造データに特化しているため，これらの複合的な構造データを直接扱うことは不可能である．その一方で，これらのデータは今後益々の増大が予想され，現実的な応用における，より踏み込んだマイニングを実現する手法として，複合構造データを包括的に扱うことの出来る柔軟かつ高精度なマイニング手法の確立は急務である．

２．研究の目的

従来の構造データを対象としたパタン（部

分構造）発見手法では，一般に，抽出されるパタンの数が膨大となる．従って，マイニング結果をより有効なものにするためには，その中からある基準で代表的または特徴的なパタンのみを高速に列挙することが求められる．これに加え，利用者にとって興味深くまた理解容易な結果を得るためには，求めるべきパタンに対する制約など，利用者が持つ領域知識の利用が重要となる．

本研究の対象である複合構造データは，従来の SDM が対象としているデータと比較してより複雑であるため，上記の要求はより本質的である．そこで本研究では，既存の SDM 手法を拡張し，(a)制約を考慮した代表的パタンの発見手法の開発，及び(b)階層的構造データに特化した特徴的パタンの定義とその効率的な発見手法の開発を目的とする．

一方，本研究の目的達成のためには，既存の SDM 手法の効果的な援用が重要となる．すなわち，既存手法の本質的な部分を，相互関係・再利用性の観点からの整理し，それらを有機的に連動させる必要がある．このことは，オントロジによる既存手法の積極的再利用に基づく高度マイニング手法構築の，基本的な方法論を与えることにもつながると考えられる．これらの展開の第一歩として，本研究では，特に研究が進んでいる木構造や系列などを対象に，既存 SDM 手法が，階層性の面からの最適化を伴う形で統合・連結・再利用できることを確認するとともに，階層構造データに対する効率的な手法の構成方法の基本的な考えや枠組みを明らかにすることを目的とする．

３．研究の方法

本研究の目的達成のため，(1)制約付き特徴的パタン発見手法の開発，(2)複合構造グラフを対象とした特徴的パタン発見手法の開発，(3)グラフ系列を対象とした特徴的パタン手法の開発の 3 点に対して研究を行った．

(1)制約付き特徴的パタン発見手法の開発

代表的な構造データの一つである順序木データを対象に研究を行った．これまでに，順序木データを対象とした特徴的パタン発見手法として飽和順序木発見手法が提案されている．本研究では，この手法を基に，木の大きさや高さなど，木の形状に関する制約の導入を行うことで，制約付きの特徴的パタン発見を実現した．また，同様の考え方を用い，別の観点からの特徴的パタンとして，制約付きの極小パタンの発見手法についても検討を行った．

(2)複合構造グラフを対象とした特徴的パタン発見手法の開発

複合構造グラフとは，頂点に複数の構造データを持つグラフであり，代表的な階層的構造データの一つである．本研究では，既存のグラフマイニング手法を中心に，各種 SDM 手法を有機的に連動させることで，複合構造グラフを対象とした種々の特徴的パタン（頻出パタン，制約付きパタン，強相関パタン）の発見手法を開発した．

(3)グラフ系列を対象とした特徴的パタン発見手法の開発

違う種類の階層的構造データとして，グラフ系列を対象としたパタン発見手法の開発を行った．系列データマイニング手法とグラフマイニング手法を援用し，かつパタン間の階層関係を考慮することで，単一の長大なグラフ系列からの，特徴的部分グラフ系列（飽和強相関部分グラフ系列）の発見手法を開発した．

４．研究成果

本研究を通じ，（複合的な）構造データを対象とした種々の特徴的パタン発見手法を新たに開発し，その成果を国際会議や学術論文誌にて発表した．以下に主な成果を示す．

(1)制約付き飽和・極小順序木パタンの発見

先述したとおり，単純な頻出パタン発見では，大量のパタンが発見されてしまうという問題が指摘されている．この問題に対して (a) 飽和パタンに代表される頻出パタンの代表元のみを発見する，(b) 利用者により与えられる制約を満たすパタンのみを発見するなどのアプローチが提案されている．両アプローチは，結果として発見されるパタン数を減少させるという意味では同じであるが，後者は制約を通じて積極的に求めるべきパタンを限定しているのに対し，前者は同じ内容の圧縮表現を求めており，その目的は全く異なる．従って，両アプローチを統合することにより，より効果的なパタン発見が実現されることが期待できる．この考えに基づき，近年，アイテム集合を対象とした両者の統合アプローチが提案されている．

本研究では，構造データに対する統合アプローチの一つとして，順序木を対象に，節点数や高さ上限などのパタンの形状に関する制約を導入することにより求めるべきパタンを限定し，その上での圧縮表現を求めるという統合アプローチについて検討し，その結果として，逆単調制約付き頻出飽和順序木の発見を実現した．

より具体的には，出現マッチと呼ばれる関係と 制約付き飽和順序木の列挙との関係に対する考察を通じ，3 種の制約付き頻出飽和順序木発見アルゴリズム RCLOCOT 及び

posCLOCOT, negCLOCOT を開発した．RCLOCOT は，制約なしの飽和順序木発見手法の素直な拡張であり，制約に関する位置限定出現マッチに基づく枝刈りを採用している．一方，posCLOCOT 及び negCLOCOT は，順序限定出現マッチと境界パタンに基づく枝刈りを用いることで，制約付き飽和順序木の効率的な発見を実現している．またすべてのアルゴリズムは，後処理としてではなく，枝刈りを伴う探索の過程で解の発見を行うという特徴を持つ．また，これら３種のアルゴリズムの有用性は，合成データ及び実データを用いた評価実験を通じて確認された．

一方，飽和パタンなどの代表元を求める手法の問題点として，ノイズに対する脆弱性が指摘されている．加えて，得られたパタンをその後の分析に利用することを考えた場合，MDL 原理などの観点から（飽和パタンなどの極大元ではなく）極小元を用いる方が適切であるという報告もなされている．これらの事から，構造データを対象とした頻出パタン発見において，ノイズを考慮した極小元の発見は一つの重要な課題であると考えられる．

これらのことを背景に，順序木データベースを対象とした新たな統合アプローチとして，(1)利用者により与えられる単調・逆単調制約を満たすパタンのうち，(2)誤差を考慮した同値類の極小元のみを高速に獲得する一連のアルゴリズムの開発を行った．すなわち，利用者が持つ対象に対する知識に基づき，例えばパタンが満たすべき節点数の下限（単調制約）や高さの上限（逆単調制約）などを用いてパタンの形状を限定し，その上で，ほぼ同じデータ集合に出現するパタン集合を一つにまとめるという，圧縮表現を求める技術を開発した．

具体的には，(1) 支持度の差に着目した制約付き δ-フリー順序木パタン（δ-FCost）と(2) 支持度の比に着目した制約付き Δ-トレランス順序木パタン（Δ-TCost）の２種のパタンを新たに考案した．さらにこれらのパタンを効率的に発見するため，制約付き δ-出現マッチングを提案するとともに，これを探索戦略の異なる既存の頻出順序木発見手法に組み込むことで，3 種の δ-FCost 発見アルゴリズムを開発した．さらに，δ-出現マッチングと Δ-TCost の関係に着目し，δ-FCost 発見アルゴリズムの一部を改変することで，Δ-TCost の発見を実現した．

構造データマイニング分野において，これまで制約付き飽和パタンを扱った研究は無い．また，ノイズを考慮した手法，及び単調・逆単調の両制約を考慮した手法は，これまで提案されていない．加えて，提案した各種手法は，無順序木や，グラフパタンの発見などへの応用も期待される．これらの事から，今回開発を行った手法群の新規性・有用性は非常に高いと考えられる．

(2)特徴的複合グラフパタンの発見

本研究では，各頂点にアイテム集合や系列などの構造データの集合を持つ複雑なグラフデータベース，すなわち複合構造グラフデータベースを対象とした頻出パタン発見手法について検討を行い，頻出部分複合構造グラフ発見アルゴリズム FMG を開発した．FMG は，グラフ構造列挙手法を用いた（a）外部構造（グラフ構造）の列挙と，集合や系列列挙手法などを用いた（b）内部構造（頂点構造)の列挙を組み合わせることで，複合構造データベースにおける頻出パタンの完全な列挙を実現している．

一方，得られるパタン数の増大という頻出パタン発見の欠点に対処するため，FMG を拡張し，パタン内の頂点を，利用者による制約を満たす代表的なパタンに限定する手法 CCFMG を開発した．CCFMG では，(a)内部単調制約による枝刈りと(b)内部飽和性による枝刈りが用いられているが，これらは階層的構造データの特徴を利用したものであり，階層的構造データを対象とした制約付きパタンマイニングの一つの基礎を与えるものとなっている．

またパタン数の増大を抑制するための別の方向への拡張として，パタンの構成要素間に強い依存性が存在するもののみを発見する手法 HFMG を新たに提案した．HFMG では，(a)頂点と頂点内のデータ間の依存性である内部相互依存性，(b)グラフ構造と頂点間の依存性である外部相互依存性の両者において強い依存性を持つパタン，すなわち強相関パタンのみを効率的に抽出することが可能となっている．

開発した一連の手法を用い，生物の代謝経路を対象とした実験を行ったところ，代謝に関わる化合物や酵素のもつ情報（アミノ酸配列，酵素番号など）を頂点にもつパタンの発見に成功するなど，応用面においても，ある程度の成果が得られている．

これまでに，複合構造データを対象としたパタン発見手法は提案されておらず，その観点で，本研究の新規性は高いと考えられる．加えて，複合構造グラフは今後ますますの増大が予想され，それらを扱うことのできる柔軟な手法や枠組みの基礎として，本研究の果たした役割は大きいと考えている．

(4)強相関飽和部分グラフ系列の発見

本研究では，違う種類の階層的構造データとして，グラフ系列を対象としたパタン発見手法の開発を行った．なお本研究は，グラフ系列としてモデル化される事象に対し，変化

の傾向を把握するとともにその要因を分析し，また，その後の変化を予測するための基礎的な材料を与えるため，グラフ系列中に繰り返し現れる意味のあるパタンを抽出することを目的としている．

単純なパタン発見では，意味のないパタンが大量に発見されることが容易に想像される．そこで本研究では，(1)相互依存性，及び(2)飽和性の観点から不要なパタンを排除し，その上で，一本の長大なグラフ系列より，頻繁に出現する部分グラフ系列を抽出するアルゴリズム CHPSS を開発した．ここで相互依存性とは，パタン全体(部分グラフ系列)とその構成要素（各部分グラフ）との依存関係を意味する．相互依存性が低いパタンは，無関係もしくは不要な構成要素を持つパタンであり，そのようなパタンを排除することで，理解しやすく，また意味のあるパタンのみが抽出されることが期待される．一方，飽和性とは同値類における代表元を表すものであり，同じものを説明するパタンの集合（同値類）から最も特殊なパタンのみを抽出することで，同じ内容を保ったまま，生成されるパタン数の削減が達成される．

これまでに，グラフ系列を対象としたパタン発見手法がいくつか提案されているが，パタン発見に頻度以外の基準を用いたものは，多くない．これに対し CHPSS では，相互依存性と飽和性という２つの代表性・特徴性を同時に用いており，新規性のみならず，有用性や技術的な貢献という意味でも，十分に評価できるものであると考えている．

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）

〔雑誌論文〕（計　５　件）

[1] 尾崎 知伸，大川 剛直，拡張出現マッチングを用いた制約付きノイズ許容極小順序木パターンの発見，情報処理学会論文誌：データベース，Vol.1，No.3，pp.20-35，2008，査読有

[2] 尾崎 知伸，大川 剛直，グラフデータベースからの頻出相互関連部分グラフ集合の発見，人工知能学会論文誌，Vol.23，No.6，pp.514-525，2008，査読有

[3] 山本 翼，尾崎 知伸，大川 剛直，構造データ集合からなるグラフデータベースからの頻出パターン発見，情報処理学会論文誌：データベース，Vol.1，No.1，pp.26-35，2008，査読有

[4] 尾崎 知伸，大川 剛直，限定的出現マッチを利用した逆単調制約付き頻出飽和順序木の発見，人工知能学会論文誌，Vol.23，No.2，pp.58-67，2008，査読有

[5] 尾崎 知伸，大川 剛直，制限付き最右拡張を用いた効率的な飽和順序木の発見，情報処理学会論文誌：データベース，Vol.48，No.SIG11(TOD34)，pp.118-127，2007，査読有

〔学会発表〕（計　８　件）

[1] Tsubasa Yamamoto，Tomonobu Ozaki，Takenao Ohkawa，Discovery of Internal and External Hyperclique Patterns in Complex Graph Databases，The 4th International Workshop on Mining Complex Data (MCD 2008) in conjunction with IEEE ICDM 2008，pp.301-309，15 December 2008，Pisa Italy

[2] Tomonobu Ozaki，Takenao Ohkawa，Mining Correlated Pairs of Patterns in Multidimensional Structured Databases，The 4th International Workshop on Mining Complex Data (MCD 2008) in conjunction with IEEE ICDM 2008，pp.275-282，15 December 2008，Pisa Italy

[3] Tomonobu Ozaki，Takenao Ohkawa，Discovery of Closed Hyperclique Patterns in a Sequence of Graphs，The Third International Workshop on Data-Mining and Statistical Science (DMSS2008)，pp.12-15，25 September 2008，Tokyo Japan

[4] Tomonobu Ozaki，Takenao Ohkawa，Mining Mutually Dependent Ordered Subtrees in Tree Databases，First International Workshop on Algorithms for Large-Scale Information Processing in Knowledge Discovery (ALSIP 2008)，pp.78-89，20 May 2008，Osaka Japan

[5] Tomonobu Ozaki，Takenao Ohkawa，Mining Correlated Subgraphs in Graph Databases，The 12th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2008)，pp.272-283，23 May 2008，Osaka Japan

[6] Tsubasa Yamamoto，Tomonobu Ozaki，Takenao Ohkawa，Discovery of frequent graph patterns that consist of the vertices with the complex structures，The Third International Workshop on Mining Complex Data (MCD'07)，pp.71-82，17 September 2007，Warsaw Poland

[7] Tomonobu Ozaki，Takenao Ohkawa，Mining Frequent Delta-Free Induced Ordered Subtrees in Tree-structured Databases，The 5th Workshop on Learning with Logics and Logics for Learning，pp.3-9，18 June 2007，Miyazaki Japan

[8] Tomonobu Ozaki, Takenao Ohkawa, Efficiently Mining Closed Constrained Frequent Ordered Subtrees by using Border Information, The 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2007), pp.745-752, 23 May 2007, Nanjing China

６．研究組織

(1)研究代表者

尾崎 知伸 (OZAKI TOMONOBU)
神戸大学・自然科学系先端融合研究環重点研究部・助教
研究者番号：40365458

(2)研究分担者

(3)連携研究者