様式 C-19

科学研究費補助金研究成果報告書

平成 23 年　4 月　21 日現在

機関番号：１４４０１
研究種目：若手研究(B)
研究期間：2009 ～ 2010
課題番号：21700168
研究課題名（和文）複合構造データベースからの関連パターン発見手法の開発

研究課題名（英文）　Correlation Pattern Mining from Complex Structured Databases

研究代表者
尾崎 知伸　（ OZAKI TOMONOBU ）
大阪大学・サイバーメディアセンター・特任講師
研究者番号：40365458

研究成果の概要（和文）：
本研究では、構造データの組み合わせである複合構造データからの知識発見に関し、(1)複合構造グラフや(2)グラフ系列、(3)多次元構造データ、(4)定量的区間系列などを対象に、特徴的な共通パターンを発見するための効率的な手法の開発を行った。またこれらの研究を通じ、従来技術の問題点である質の低いパターンの生成を抑制し、有意義で特徴的かつ解釈容易なパターンを効率的に発見するための基礎的な方法論について一定の知見を得た。

研究成果の概要（英文）：
In this research, several algorithms were developed mining common characteristic patterns from complex structures such as (1)graphs with composite properties, (2)graph sequences, (3)multi-dimensional structured databases and (4)quantitative event sequences. Through the research, we obtained a basic and general framework which suppresses meaningless low quality patterns and finds only significant and comprehensive patterns.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	1,700,000	510,000	2,210,000
2010 年度	1,600,000	480,000	2,080,000
年度			
年度			
年度			
総　計	3,300,000	990,000	4,290,000

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：データマイニング，構造データ，関連パターン，アルゴリズム

１．研究開始当初の背景

グラフ構造データは、多様な関係を自然に表現することが可能であり、種々の分野で広く利用されている。また、対象のより精細なモデル化においては、(i)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」や、(ii)ある構造の中に別の構造が現れる「階層的構造データ」など、構造データの組み合わせである「複合構造データ」が必要となると考えられる。

構造データ及び複合構造データは、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データ分析のツールとして、これらの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。

一方、一般的に用いられている頻度に基づくデータマイニング（パターンマイニング）

手法では、当たり前のパターンや偶発的なパターン、解釈困難なパターンなど、「質の悪いパターンが大量に生成される」という問題が指摘されており、実応用における大きな障壁となっている。

複合構造データマイニングに対し、要素間の関係性を強く考慮する関連パターン発見技術を導入することは、(i)質の悪いパターンを排除するとともに、(ii)得られたパターンの解釈容易性を向上させるという意味で、パターンマイニングの質の向上に有効であるが、その有効性・重要性・緊急性に関わらず、国内外を問わず、ほとんどその研究が行われていないというのが現状である。

これらのことを背景に、本研究では、関連パターン発見技術と複合構造データマイニング技術を再度見直し、より広範囲な複合構造データを対象とした柔軟かつ表現力豊かな関連パターン発見の実現を目指した.

２．研究の目的

本研究では、幅広い応用を念頭に、主に(i)適用対象の拡大、及び(ii)関連性評価の柔軟化の観点から、複合構造データに対する関連パターン発見手法の高度化及び汎用化を目的とする。

適用対象の拡大には、既存の複合構造データマイニング技術の効果的な援用が重要となる。本研究では、既存の関連パターン発見及び複合構造データマイニング技術を詳細に検討し、両者の長所を損なわない形での有機的な統合・連動を実現することを目的とする。またこれにより、領域や対象に依存しない、より汎用的な複合構造データに対する関連パターン発見技術の基礎を確立することを目指す。

一方、関連性評価の柔軟化に関しては、関連性をどのように定義するのか、という関連パターン発見の本質的な部分を検討する。既存手法では、「パターン間やパターン内で、すべての要素が互いに共起している」という基準を採用している。これを基本に、本研究では、より制御しやすく、また得られる結果が解釈しやすい基準の開発を目的とする。またこのことを通じ、複合構造データに限らず、広く一般に適用可能な関連パターンに関する新たな基準の開発を目指す。

３．研究の方法

本研究の目的達成のため、(1)複合構造データマイニングの柔軟化、及び(2)多次元構造データを対象とした関連パターン発見手法の開発、(3)グラフ系列を対象とした関連パターン発見手法の開発の３点に対し研究を行った。

（１）複合構造データマイニングの柔軟化

より広範囲の構造データを対象としたデータマイニング手法を実現するため、(i)外部及び内部重み付きグラフデータベース、(ii)数値属性集合付きグラフ、(iii)定量的区間イベント系列のそれぞれを対象に、特徴的パターン発見手法の開発を行った。

これらのデータは、それぞれ一部に定量的データ（数値データ）を含むものであり、主に記号データを対象としている（複合）構造データマイニング手法に対し、その適用範囲を本質的に拡張するものであると考えている。

（２）多次元構造データを対象とした関連パターン手法の開発

グラフデータを対象とした関連パターン発見アルゴリズム HSG を多次元構造データへと拡張するとともに、より特徴的なパターン集合のみを抽出するために、飽和性（パターンの大きさに関する極大性）を満たすパターン集合のみを発見するアルゴリズムを開発した。またより効率的な発見を実現するため、頻度、関連性、飽和性のそれぞれに着目した最適化手法を開発した。

（３）グラフ系列を対象とした関連パターン発見手法の開発

単一のグラフ系列に対する関連パターンとして、頻出飽和関連部分グラフ系列(FCSS)を提案した。また、パターンに対する関連性基準の柔軟化の一つとして、(i)系列パターンの各構成要素（部分グラフ系列）の大きさ、(ii)互いに関連しあう構成要素数、(iii)関連の強さを表す相関係数の３つのパラメタからなる関連性基準を提案すると共に、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た効率的な FCSS 発見手法を開発した。

加えて，動的グラフを対象に，時間差に着目したパターン発見及びパターン集合の集約に関しても検討を行った。

４．研究成果

本研究を通じ、種々の複合構造データを対象とした効率的な特徴的パターン発見アルゴリズムを開発した。またそれら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）重み付きグラフデータベースからの特徴的パターンの発見

本研究では、グラフマイニングの一つの発展として、グラフ自身及びグラフの各構成要素に対し、その重要性や信頼性、意義などを表す重みが付与された、外部及び内部の重み付きグラフデータベースからのパターン発見について検討を行った。

外部及び内部重みの双方を考慮したパタ

ーンに対する新たな評価基準として、(i)各データグラフ中における複数の出現の集約方法、(ii)外部及び内部重みそれぞれの観点からの評価の統合方法の2側面から考察を行い、それらの組み合わせとして 10 数種の評価基準を考案した。また、正の重みに加え、負の重みを制約として利用する枠組みについても検討を行った。本研究では、これらの基準をもとに、頻出パターンの代表元である飽和パターン及び極大パターンを発見する一般的なアルゴリズム wgMiner を開発した。wgMiner は、部分グラフマイニングに関する種々の技術を統合したものであり、評価基準の上界値に基づく枝刈りに加え、飽和性に基づく枝刈りを採用している。加えて、飽和性・極大性チェックのためのコストを軽減するため、逆探索の技術により構築されるパターン空間を、行きがけ順（pre-order traversal）に辿りながらパターンを生成すると共に、帰りがけ順（post-order traversal)にその飽和性・極大性をチェックするという戦略を採用している。

外部及び内部重みを同時に考慮する枠組みとして、アイテム集合発見（バスケット分析）分野において、ユーティリティに基づくアイテム集合発見と呼ばれる枠組みが提案されている。本研究は、ユーティリティに基づくアイテム集合発見のアイディアを構造データへと拡張・適用するに留まらず、扱える重み範囲の拡大、評価関数の柔軟化、及び特徴的パターン（飽和パターン・極大パターン）への限定という本質的な拡張を伴うものであり、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評価できるものであると考えている。

（2）数値属性集合付きグラフからの頻出パターン発見

グラフマイニングの研究の多くは、単純なラベル付きグラフを対象としているが、対象のより自然かつ精密な表現を考えた場合、必ずしもラベル付きグラフで十分であるとは限らない。そこで本研究では、対象のより精密な表現手段として、頂点（辺）に数値属性の集合が付与された頂点（辺）数値属性付きグラフを採用し、そこからのパターン発見について議論を行うと共に、数値属性付きグラフから頻出パターンを発見するアルゴリズム FAG-gSpan を開発した。FAG-gSpan は、頻出部分グラフを発見するアルゴリズム gSpan による(i)ラベル付きグラフパターンの列挙と、高密度クラスタに基づいて定量的アイテム集合を発見するアルゴリズム QFIMiner を利用した(ii)定量的アイテム集合の割り当てを繰り返すことで、特徴的な部分グラフを発見する。また抽出された数値属性付きグラフパターンに対し、出現をもとに(i)支持度の計算と(ii)標準形判定を導入することにより、重複なく頻出パターンを列挙することが可能である。

前述の通り、FAG-gSpan は、gSpan と QFIMiner の組み合わせにより構成されているが、その枠組み自体は、任意のグラフ構造列挙アルゴリズム及び定量的アイテム集合列挙アルゴリズムを利用することが可能である。その意味で、今回開発した枠組みは、一般的かつ柔軟なものであると考えている。

（3）定量的区間イベント系列からの頻出パターン発見と分類への応用

本研究では、時間幅を持つイベント系列の集合である定量的区間イベント系列データベースを対象としたパターンとして、(i) 事象の区間長や事象が発生する時間差といった定量的情報、及び (ii) 事象が観測されなかったという否定的情報の2つの情報を考慮した正負定量的区間パターンを提案した。さらに、このパターンを効率的に列挙する手法として、特に負リテラル（否定情報）の生成法に着目し、(i)リテラル法、(ii)動的法及び(iii)後処理法と呼ばれる 3 種の手法を考案した。加えて、開発した手法を、実データを含む数種の分類問題における特徴生成へと応用し、その有用性を評価・確認した。

定量的情報及び否定情報を同時に考慮する手法はこれまで提案されておらず、その観点で、本研究の技術的な新規性は高いと考えている。加えて、特に応用面で、その適用範囲の拡大が期待できると考えている。

（4）多次元構造データからの関連パターン発見

本研究では、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''を対象に、属性間に跨る特徴的関連パターンを発見する枠組みについて研究を行い、飽和性を満たすパターン集合のみを効率的に発見するアルゴリズム CHPMS を開発した。ここで、属性間に跨るパターンとは、異なる次元におけるパターンの組み合わせ（集合）を意味する。CHPMS では、互いに強く関連しあう、異なる次元からそれぞれ得られるパターンの集合を、最終的に獲得すべきパターンと考えている。また、関連性の基準として、h-確信度 (h-confidence)を採用している。

CHPMS で得られるパターンは、アイテム集合発見やグラフパターン発見におけるハイパークリークパターンの概念を多次元構造データベースへと拡張・適用したものである。従って、獲得される各パターン集合は、お互いがお互いを説明しあうパターンの集合と捉えることができ、ある次元から得られるパターンに対して、同一集合中の他の次元から得られるパターンを用いた解釈や意味づけ

を可能にするものである。従って、この様なパターン集合を発見することは、次元を跨るパターン間の関係を把握することが期待できると共に、対象データベースにおける新たな発見やより深い対象の理解につながると考えられる。

本研究では、より特徴的なパターン集合のみを抽出するために、獲得すべきパターンを飽和パターン、すなわち、同一頻度及び同一h-確信度を持つパターン集合の中で構造的な意味で極大なパターン集合に限定している。CHPMS では、より効率的に飽和パターンを導出するため、(i)次元毎に既存の飽和パターン発見アルゴリズムを適用すると共に、得られる飽和パターンを、その一般性（generality-ordering）に基づき木構造で管理し、(ii)それら複数の木構造を相互に走査することでパターン集合を獲得するという2段階のアルゴリズムを採用している。アルゴリズムを2段階に分けることにより、(i)飽和パターン発見においては、今後開発されるであろう各種構造データに対するより効率的なアルゴリズムの柔軟な組み込みが可能となっている。また(ii)パターン集合獲得では、木構造で表現されるパターン間の関係を積極的に利用することで、効率的なアルゴリズムに必要不可欠な枝刈りが、(a)頻度、(b)h-確信度、(c)飽和性の3つの観点から統一的かつ自然な形で実現されている。

開発した手法は，パターンを集合として獲得することで集合内の各パターンの解釈を助けるものであり，その点で，パターンマイニングの質の向上に直接つながるものであると考えている.

（５）グラフ系列からの関連パターン発見

本研究では、単一のグラフ系列を対象とした関連パターンとして、頻出飽和関連部分グラフ系列(FCSS)を提案すると共に、効率的なFCSS 発見手法 CorSSS を開発した。

一般に、構造データに対するパターンの関連性基準は、パターン全体とその構成要素（基本要素）との関連性に基づく。また多くの既存研究では、「関連性がある」と判断する基準として、基本要素を頂点、その関連性を辺とする 関連性グラフがクリークであることを要請する。しかし、この要請は非常に厳しくまた画一的で柔軟性に欠けるものであり、特に応用面で、必ずしも有効であるとは限らない。そこで本研究では、パターンであるグラフ系列に対し、その（連続する）部分系列をパターンを構成する基本要素と考え、(a)各基本要素の大きさ（長さ）m、(b)互いに関連しあう構成要素数ｋ、(c)関連の強さを表す相関係数θの3つのパラメタからなる関連性基準（m、θ、ｋ)-関連性を提案した。この基準は、各構成要素を頂点、それらの関連性を辺とする関連性グラフにおける K-plex の概念に相当するものであり、従来研究で用いられるクリークに基づく関連性基準の一種の緩和となっている。

本研究では、提案した関連性に基づき、頻出かつ飽和な部分グラフ系列(FCSS)を発見するためのアルゴリズム CorSSS を開発した。CorSSS は、系列パターン発見の代表的なアルゴリズムである GSP の概念と、研究代表者らがこれまでに開発した(i)木構造によるパターンの管理及び(ii)複数の木構造走査によるパターンの列挙技術とを連動させることで、効率的な FCSS 発見を達成している。

これまでにグラフ系列を対象としたパターン発見技術がいくつか提案されているが、関連性を考慮した手法はなく、その点で CorSSS の新規性は高いと考えられる。また、利用者により制御可能な関連性基準を導入することで、より柔軟に偶発的なパターンの生成を回避することが期待できる。これらのことより、CorSSS は、「質の悪いパターンの抑制」という面でパターンマイニングの質の向上につながる成果であると考えている。

（６）動的グラフからの特徴的パターン発見

動的グラフ（グラフ系列）からの特徴的パターンとして、粒度の異なる3種の変化パターンを定義すると共に、それらの効率的な列挙手法の開発を行った。また、(i)パターンの構造及び説明範囲の類似性に着目したパターン選択技術の開発、及び(ii)パターン集合の構造化（ネットワーク化）技術を新たに開発することで、(従来手法で用いられる)頻度とは異なる基準による、特徴的なパターンの絞込み及びパターンのランキングを実現した。

開発した手法は、関連性の観点から、得られるパターン集合を構造化するものであり、パターン集合内における各パターンの役割や位置づけを理解するのに有用であると考えている。

以上示したように、本研究では、種々の構造データ及び複合構造データを対象に、特徴的パターン発見のための効率的なアルゴリズムの開発を行なった。これらの研究を通じ、従来技術の問題点である質の低いパターンの生成を抑制し、有意義で特徴的かつ解釈容易なパターンを効率的に発見するための基礎的な方法論について一定の知見が得られたと考えている。

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　3件)

[1] 三好裕樹, 尾崎知伸, 江口浩二, 大川剛直, 定量的アイテム集合付き単一グラフからの頻出パターンマイニング, 人工知能学会論文誌, 査読有, Vol.26 No.1, (2011), 284-296

[2] 信田正樹, 尾崎知伸, 大川剛直, 内部および外部重みを考慮した頻出部分グラフマイニング, 情報処理学会論文誌：データベース,査読有,Vol.3 No.2,(2010), 1-12

[3] 山本翼,三好裕樹,尾崎知伸,大川剛直, 複合構造グラフからの頻出強相関パターン発見, 情報処理学会論文誌：データベース, 査読有, Vol.2 No.3, (2009), 53-66

〔学会発表〕(計　9件)

[1] Tomonobu Ozaki and Minoru Etoh,Closed and Maximal Subgraph Mining in Internally and Externally Weighted Graph Databases, The 4th International Symposium on Mining and Web, 2011.3.22, Biopolis（Singapore）

[2] Yuuki Miyoshi,Tomonobu Ozaki and Takenao Ohkawa,Mining Interesting Patterns and Rules in a Time-evolving Graph, The International Multi Conference on Engineers and Computer Scientists 2011, 2011.3.28, Hong Kong (China)

[3] 信田正樹, 尾崎知伸, 大川 剛直, 外部・内部重み付きグラフマイニングにおける評価尺度の比較, 人工知能学会 第 77 回 人工知能基本問題研究会, pp.31-36, 2010.3.17, 札幌

[4] 柏木潔, 尾崎知伸, 大川 剛直, 被覆集合に着目したグラフデータベースからの分割パターンの発見, 人工知能学会 第 77 回 人工知能基本問題研究会, pp.19-24, 2010.3.17, 札幌

[5] Tomonobu Ozaki and Takenao Ohkawa, Efficient Discovery of Closed Hyperclique Patterns in Multidimensional Structured Databases, The 3rd International Workshop on Mining Multiple Information Sources, pp.533-538, 2009.12.6, Miami (USA)

[6] Yuuki Miyoshi、Tomonobu Ozaki and Takenao Ohkawa, Frequent Pattern Discovery from a Single Graph with Quantitative Itemsets, The 3rd International Workshop on Mining Multiple Information Sources, pp.527-532, 2009.12.6, Miami (USA)

[7] Fumiya Nakagaito, Tomonobu Ozaki and Takenao Ohkawa, Discovery of Quantitative Sequential Patterns from Event Sequences, The 2009 International Workshop on Domain Driven Data Mining, pp.31-36, 2009.12.6, Miami (USA)

[8] Masaki Shinoda, Tomonobu Ozaki and Takenao Ohkawa, Weighted Frequent Subgraph Mining in Weighted Graph Databases, The 2009 International Workshop on Domain Driven Data Mining, pp.58-63, 2009.12.6, Miami (USA)

[9] Tomonobu Ozaki and Takenao Ohkawa, Discovery of Correlated Sequential Subgraphs from a Sequence of Graphs, The 5th International Conference on Advanced Data Mining and Applications, pp.265-276, 2009.8.18, Beijing (China)

６．研究組織

(1)研究代表者

尾崎　知伸　（ OZAKI TOMONOBU ）

大阪大学・サイバーメディアセンター・特任講師

研究者番号：40365458