様式Ｆ－１９

科学研究費助成事業（学術研究助成基金助成金）研究成果報告書

平成２５年　５月２１日現在

機関番号：32665
研究種目：若手研究（Ｂ）
研究期間：2011～2012
課題番号：23700173
研究課題名（和文）
多様な関連性に着目した特徴的部分構造パターンの発見に関する研究
研究課題名（英文）
Mining Interesting Substructures based on Various Relevance Measures
研究代表者
尾崎　知伸（ OZAKI TOMONOBU ）
日本大学・文理学部・准教授
研究者番号：40365458

研究成果の概要（和文）：
本研究では、時間と共に構造が変化する動的ネットワークを対象に、ベースとなる連結グラフパターンとそこから拡張されるリンクの対からなるリンク生成パターンを想定した上で、複数のリンク生成パターンを集約することで得られるより特徴的な（メタ）パターンとして、関連・対比パターンを新たに提案するとともに、その効率的な発見技術を開発した。また、２つの実データを用いた実験により、開発した各手法の有効性を確認した。

研究成果の概要（英文）：
In this research, by upgrading hyperclique patterns and conditional contrast patterns in itemset mining into graph domain, we propose correlation and contrast link formation patterns in a dynamic network, respectively. Discovery of sets of link formation patterns having opposite characteristics, i.e., correlation and contrast, can expect to obtain deep understanding of target dataset. Experiments using real world datasets confirm the effectiveness of the proposed framework.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
交付決定額	2,900,000	870,000	3,770,000

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

１．研究開始当初の背景

構造データマイニングとは、蓄積された大量のグラフなどの構造データから有益な知識・知見を得るための技術の総称であり、近年、種々の観点から積極的な研究が行われている。また、構造データマイニングの一つの発展である、付加的な情報を伴う構造データを対象とした複合構造データマイニングに関しても、その研究が進められている。

研究代表者もこれまで、特にデータ表現の観点から構造データマイニングを拡張した複合構造データマイニングの一つとして、(a)多様な属性を伴うノードから構成されるネットワーク（例：化合物の詳細情報を含んだ代謝ネットワーク）や、(b)時間的変化を伴うネットワーク（例：社会ネットワークとその変化）に関するパターンマイニング手法を提案している。なお、パターンマイニングとは、データ中に存在する特徴的な組み合わせ（部分構造）を効率的・効果的に抽出する技術の総称である。

一般にパターンマイニングに対しては、自明・偶発的・理解困難という意味で、低品質なパターンが大量に生成されるという本質的な問題が従来から指摘されており、これまでにその解決策として、(i)パターンに対する

制約の導入や(ii)主観的尺度の利用、(iii)代表元への限定や(iv)パターン集合への集約など、様々なアプローチが提案されている。また、特に構造データを対象としては、(i)強く関連するパターンの集合である関連構造パターン集合や、(ii)関連しあう部分構造だけから構成される関連複合パターンなど、主に理解容易性の向上を目的とした、関連性に基づく新たな特徴的部分構造パターンが提案されている（図１参照）。

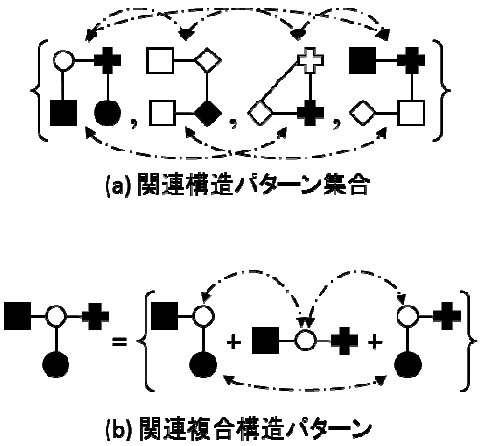

図 1:構造データに対する関連パターン

研究代表者は、「低品質・理解困難なパターンの大量生成」というパターンマイニングの本質的な問題を解決する一つのアプローチとして、複数の視点や文脈を用いたパターンの考察及び相互関連付け、すなわち、「各パターンに対する類義語や反義語に相当するパターンの抽出とその付与」が有望であると考えている。また、類義語や反義語はといった文脈の一部は、関連構造パターンや関連複合構造パターンとして与えられると期待している。このように、複数の観点から各パターンに対して関連情報を付与することは、それ自体パターンの理解容易性の向上につながると考えられる。加えて、付与された関連情報を考察することで、パターン集合全体における各パターンの意味や役割、その適用範囲を考慮した重要性の判断に対しても有効であると考えられる。その一方で、現状の構造データマイニング技術を考えた場合、扱える関連性の種類が乏しいことに加え、複数種の関連性を同時に考慮したパターンに対する集約技術は提案されていないなど、必ずしも十分な問題解決が達成されているとは言えない。

これらのことを背景に、本研究では、関連性評価の多様性の観点から、新たなパターンを提案するとともに、構造データマイニング技術の適用範囲を拡大することを目指した。

２．研究の目的

本研究では、生物情報学や化学分野、及び、Web マイニングや社会ネットワーク分析などの幅広い応用を念頭に(i)関連性の多様化による構造データマイニングの適用範囲の拡大、及び(ii)効果的なパターン集約技術の開発を行う。

関連性を表すパターンの開発においては、得られる結果の理解容易性を確保するため、扱う関連性自体の直感的な意味が重要となる。例えば、関連構造パターン集合では「パターンが出現する外延集合に関連性がある」という基準を、関連複合構造パターンでは「パターンを構成する部分構造の出現位置に関連性がある」という基準をそれぞれ採用している。本研究では、正の関連性に加え、負の関連性（排他性）を加味するなど、可能な限り異なる種類の関連性を準備することが望ましいと考えている。また従来手法では、共起性に基づく基準を用いることが多いが、本研究では、オントロジー的な側面や適用範囲の拡大をも視野に入れて研究を行う。

第二の目的であるパターンの集約とは、パターン集合全体からそれらを推測することの出来る少数のパターンを導出することを表す。本研究では、パターン自体の理解容易性及び他パターンの理解へ与える影響など、「理解の観点」から各パターンを定量的に評価するとともに、パターンが持つ関連情報の多様さやパターンの適用範囲、背景知識（既得パターン）との整合性などを考慮した集約方法を開発することを目指す。また、実利用における利便性を考え、集約に関する観点の設定や集約度合いの調整等、利用者による集約の制御についても検討を行うことを予定している。

３．研究の方法

本研究の目的達成のため、「グラフデータを対象とした関連性評価基準の柔軟化と新たなパターンの提案」を中心に、以下の方法で研究を行った。

まず、既存手法を(i)対象としているデータの構造や(ii)利用している関連性の一般性・普遍性、(iii)構造データとの親和性の 3 つの観点から整理する。なお、既存手法に関する情報収集には、主要な国際会議の論文集を用いる。

次に、これまでアイテム集合発見分野で提案されている間接相関ルールやその発展である条件付き対比パターンを中心に、それら技術の構造データへの拡張・適用を行う。条件付き対比パターンとは、「形状は非常に類似するが、その説明範囲（パターンを含むデータ集合）はそれぞれ大きく異なる」という条件を満たすパターンの集合であり、形状と説明範囲という文脈で複数のパターン同士

を関連付けるものである。研究代表者は既に、条件付き対比パターンのグラフデータへの拡張に関する検討を始めており、本研究ではこれらを発展させ、複合構造データへと適用する。その際、研究代表者による関連構造パターン集合発見アルゴリズムの開発ノウハウを援用する。関連構造パターン集合とは、「形状は大きく異なるが、その説明範囲は良く似ている」パターンの集合であり、条件付き対比パターンと逆の関連性を持つ。逆の関連性を利用する２手法を準備することにより、より効果的に、構造データを対象としたパターン発見手法の適用範囲の拡大が達成されると考えている。

最後に実データを用い、開発した手法の定量的・定性的な評価を行う。データの準備には、Web 上で公開されている行動データ等を利用する。入手した行動データから、複合構造データの一種である動的ネットワークを構築し実際のパターンマイニング実験に用いる。実験では、実行効率（計算時間）などの定量的な評価に加え、開発する関連性評価基準を質的な面から評価するため、実際に得られたパターンを定性的な面からも評価する。

４．研究成果

本研究を通じ、時間とともに構造が変化する動的ネットワーク上でのリンク生成パターンに関し、関連・対比パターンを新たに提案するとともに、その効率的な発見技術を開発した。加えて、関連性評価基準の多様化や深化を目的に、主に社会ネットワーク分析分野で開発されている類似性評価技術を拡張し、履歴データからのネットワーク同定技術を開発した。また、これらの研究成果を国際会議に発表した。以下、それぞれについて説明する。

（１）リンク生成に関する関連・対比パターン発見アルゴリズムの開発

リンク生成パターンとは、動的ネットワーク上でのパターンの一つであり、ベースとなる連結グラフパターンとそこから拡張されるリンクの対からなるパターンである。またこのパターンは、ベースパターンをリンクが生成される条件と見做すことで、ネットワーク成長に関する局所的な「条件と結果の対」を表すパターンとなっている。本研究では、複数のリンク生成パターンを集約することで得られるより特徴的なメタパターンとして、関連・対比パターンを新たに提案するとともに、その効率的な発見技術を開発した。

具体的には、同一のベースパターンから、強い関連性を持ってほぼ同時に生成されるリンクの集合を、リンク生成に関する関連パターンとして獲得する技術を開発した。また、関連性評価をより柔軟に行うため、ハイパークリーク及び疑似クリークに基づく２手法を開発した。一方、関連パターンの逆の意味を持つパターンとして、同一のベースパターンから強い関連性を持って、ほぼ排他的に生成されるリンク集合をリンク生成に関する対比パターンとして獲得する技術を開発した。ここで、(i)リンク生成に関する関連パターンはネットワークにおける共進化を、逆に(ii)対比パターンは成長の分岐点を表すパターンとなっている点に注意が必要である。このように、同一データから、2 つの異なる性質を持つメタパターン（関連パターン・対比パターン）を導出することで、対象データに対するより深い理解が期待できる。また、関連性評価基準として（疑似）クリークを用いることで、導出すべきパターンに要求する「関連性」の意味をより明確にすることに成功したと考えている。

本研究で開発した各手法に対し、電子メールとモバイル通信に関する２つの実データを用いた実験により、その有効性の評価を行った。その結果、概ね現実的な計算時間で解が得られることが確認できた。また、実験を通じて実際に得られたパターンの例を以下に示す。

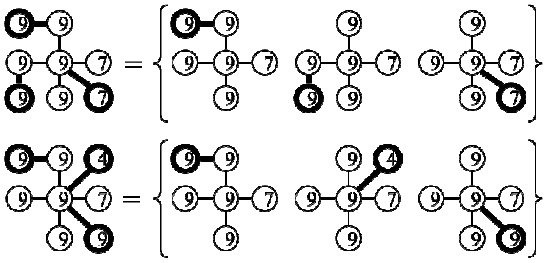

図 2：モバイル通信データセットから発見された関連パターン（上）と対比パターン（下）

リンク生成に関する関連・対比パターンの一つの拡張として、ベースパターンの生成過程に関する制約の導入についても検討を行った。これは、時間を考慮したベースパターンの細分化と捉えることも可能である。通常のリンク生成パターンでは、新しいリンクが生成される前にベースパターンが構築されていればよい。これに対し、ベースパターンを構成する各辺や頂点の生成順序をも考慮することにより、より細分化された、ある意味で精細なパターンの導出が期待できると考えている。

（２）履歴データからのネットワークの同定技術の開発

関連性評価技術の深化を目的に、履歴デー

タをネットワーク化する技術を開発した。より具体的には、ユーザの行動履歴や属性データを対象に、種々の類似性・影響力などの関連性評価技術を用いることでユーザ間のつながりの強さを推定し、ユーザ間のネットワークを構築するというものである。またこれらの技術を時間軸上にそって展開することで、動的ネットワークを獲得する方法を検討した。これにより、時間的な変換を伴う関連性評価基準に関して一定の知見を得るとともに、そのパターンマイニングへの適用に関する見通しを得た。

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５件）

[1] Tomonobu Ozaki and Minoru Etoh, Social Network Discovery from Multiple Log Data through a Behavior Model, The 2012 26th International Conference on Advanced Information Networking and Applications Workshops (WAINA 2012), 2012.03.27, Hakata (Japan)

[2] Tomonobu Ozaki and Minoru Etoh, Experimental Analysis of the Effects of Social Relations on Mobile Application Recommendation, The International MultiConference on Engineers and Computer Scientists 2012 (IMECS 2012), 2012.03.14, Hong Kong (China)

[3] Tomonobu Ozaki and Minoru Etoh, Social Network Inference of Smartphone Users based on Information Diffusion Models, The 7th International Conference on Advanced Data Mining and Applications (ADMA'11), 2011.12.18, Beijing (China)

[4] Tomonobu Ozaki and Minoru Etoh, Correlation and Contrast Link Formation Patterns in a Time Evolving Graph, ICDM 2011 Workshop on Contrast Data Mining and Applications (ContrastDM), 2011.12.11, Vancouver (Canada)

[5] Tomonobu Ozaki and Minoru Etoh, Estimation of Implicit User Influence from Proxy Logs -- An empirical study on the effects of time difference and popularity, The International Conference on Knowledge Discovery and Information Retrieval (KDIR2011), 2011.10.29, Paris (France)

６．研究組織
(1)研究代表者

尾崎　知伸（ OZAKI TOMONOBU ）
日本大学・文理学部・准教授
研究者番号：40365458

(2)研究分担者

(3)連携研究者