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：１７１０２
研究種目：基盤研究（C）
研究期間：2008～2010
課題番号：２０５０００１６
研究課題名（和文）　グラフパターン言語の計算論的学習理論とグラフマイニングへの応用
研究課題名（英文）　Machine learning theory for graph pattern languages and its applications to graph mining

研究代表者
正代　隆義（SHOUDAI TAKAYOSHI）
九州大学・システム情報科学研究院・准教授
研究者番号：５０２２６３０４

研究成果の概要（和文）：
化学化合物データやネットワークトラフィックデータといったようなグラフ構造を持つデータを対象として、そのデータに潜むグラフ構造パターンを抽出するための手法を開発した。特に、対象となるグラフ構造データのグラフ理論的な木構造的性質、例えば外平面性や木幅に着目し、一連の表現力豊かなグラフ構造パターンを設計した。我々の提案したアルゴリズムは実データに対して高速に動作し、いくつかの興味深いグラフ構造パターンを発見することに成功した。

研究成果の概要（英文）：
We proposed a series of techniques for extracting graph-structured patterns efficiently from graph-structured data, such as chemical compound data, HTML/XML data, network traffic data, and so on. In order to design expressive graph-structured patterns, we focused on tree-like properties (e.g., outerplanarity, tree-width) of target data. Our proposed algorithms run efficiently on real data, and discovered a number of interesting graph-structured patterns.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００８年度	1,300,000	390,000	1,690,000
２００９年度	1,100,000	330,000	1,430,000
２０１０年度	1,100,000	330,000	1,430,000
年度			
年度			
総　計	3,500,000	1,050,000	4,550,000

研究分野：計算機科学
科研費の分科・細目：情報学・情報学基礎
キーワード：計算論的学習理論、知識発見とデータマイニング、グラフアルゴリズム

１．研究開始当初の背景

(1) ネットワーク技術の急激な進歩にともない、ウェブページに代表されるテキストデータの利用が急速に進みつつある。とくに、安価な大容量記憶装置の発達を背景として、HTMLやXML データに代表される木構造データは、その規模を日増しに増大させている。また、近年では、化学化合物データやウェブのリンク情報といったようなグラフ構造データから情報抽出を行うことに関心が集まっている。そのようなデータベース中で、多くのグラフ構造データに共通して現れる特徴的な構造を見出すことは、データが持つ内面的性質と構造的性質を結びつけるようなルールを発見する手助けとなる。こうした応用面からの動機付けをもとに、グラフ構造データからのデータマイニング技術「グラフマイニング」が注目を浴び、現在発展を続けている。

(2) グラフマイニングの一連の研究の中で、我々は、これまでに、木構造データからのデータマイニングの理論を構築した。さらにそういった理論面の成果が机上の空論でないことの実証として、木構造パターンの多項式時間機械学習アルゴリズムによる木構造データからのデータマイニングシステムを実現してきた。

(3) さらに、木構造データへのこうした研究成果と経験に基づいて、木構造では表現しえない構造をもつデータからデータマイニング技術の開発を行い、成果をあげている。例えば、化学化合物はその多くが外平面的グラフとよばれるグラフのサブクラスで表現可能で、そのクラスは一般のグラフではNP 完全になってしまうような計算問題でも多項式時間アルゴリズムがあるという良い性質をもつ。我々は、ブロック保存型外平面的グラフパターン(Block-Preserving Outerplanar Graph Pattern、 BPOグラフパターン)と呼ばれるグラフパターンとアプリオリ法に基づくグラフマイニング手法を提案し、化学化合物データベース中に潜む新しいパターンの発見に成功した。

２．研究の目的

本研究の目的は、グラフ構造データベースに蓄積された膨大な情報から役に立つ科学的情報を抽出するための基盤となる学習理論の構築と得られた情報を幅広く提供するためのグラフマイニングシステム開発である(図1)。そのために次の2点に重点を置いた。

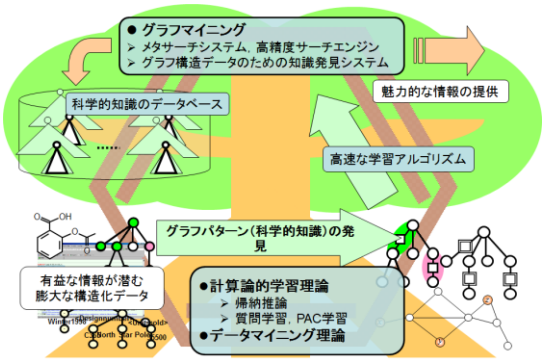

図 1 グラフパターン言語の計算論的学習理論とグラフマイニングへの応用

(1) グラフ構造データベースから知識発見のための堅固な理論的基盤を築くため、学習モデルとして正データからの多項式時間帰納推論および質問学習（厳密学習）を用い、グラフパターンのクラスの多項式時間学習可能性を明らかにする。

(2) 化学化合物や生物化学的なデータベースなどの実データをターゲットに、グラフ理論、アルゴリズムと計算量理論に基づく考察を十分に行って、現実的な時間で効果的な知識発見を行うグラフマイニング手法を開発する。

３．研究の方法

(1) 目標を達成するために、グラフデータの特徴を表わすグラフパターンとして項グラフパターンを用いる。項グラフパターンは、超辺を変数としてもち、超辺置換によるグラフ代入を行うことのできるグラフパターンである。さらに詳細な知識表現として、項グラフパターンを扱うことができるフォーマルシステム(Formal Graph System)を用いる。

(2) グラフ理論におけるグラフクラスを基に定義された項グラフパターンのクラスに対して、これまでの研究成果を基に、できるだけ一般化されたパターン照合アルゴリズムを設計する。

(3) 帰納的論理プログラミング(Inductive Logic Programming (ILP)) の手法に基づき、グラフデータからの知識発見システムの設計と実装を行う。とくに、前処理、パターン発見、検証の一連の対話的な過程による知識発見を考慮した効率化を行う。

４．研究成果

(1) 化学化合物をターゲットにしたデータマイニングアルゴリズムの設計と解析を行った。これまでに我々が提案した新しいグラフ構造パターンであるブロック保存型外平面的グラフパターン(Block-Preserving Outerplanar Graph Pattern、 BPO グラフパターン)に対する 2 つの基本的アルゴリズム、すなわち、BPO グラフパターンの多項式時間照合アルゴリズムとグラフ集合を説明する極小一般化 BPO グラフパターンの多項式時間発見アルゴリズムを与えた。その結果、BPO グラフパターンの多項式時間機械学習可能性が明らかになった。さらに、提案した学習アルゴリズムのアイデアに基づくデータマイニングアルゴリズムを提案した。このアルゴリズムを化学化合物データベースに適用し、規模耐久性および計算速度、さらにパターンの表現能力において評価を行い、いずれの評価においても十分満足すべき結果が得られた。

(2) BPO グラフパターンは、異なるブロック間の接続関係を表現するパターンであり、ブロックの内部のパターンを表現することができない。そこで、我々は、外部拡張可能外平面的グラフパターン(Externally Extensible Outerplanar Graph Pattern、 EEO グラフパターン)を導入し、多項式時間照合アルゴリズムと機械学習理論に基づくデー

タマイニングアルゴリズムを与えた。化学化合物データベース上での計算機実験の結果、EEO グラフパターンによるデータマイニング手法の現実的な処理能力と、BPO グラフパターンに対する EEO グラフパターンの表現力の優位性が示された(図 2)。

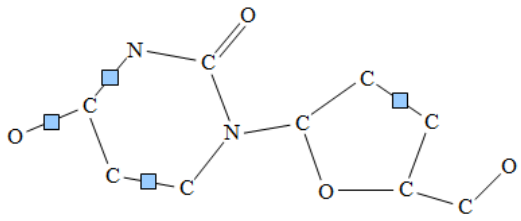

図 2 AIDS Antiviral Screen Dataset (http://cactus.nci.nih.gov/)に含まれる化学化合物 42,689 個のうち、アクティブかつ外平面的グラフで表現可能である 302 個に頻出(頻出度 0.3)な EEO グラフパターンのひとつ。

(3) グラフ構造データベースから意味のある知識を抽出するためには、まずそのデータ中に潜むパターンをどのようにグラフパターンとして表現するか、そしていかにしてうまくそのグラフパターンを発見するかが鍵となる。そこで、BPO グラフパターンの計算量理論的な性質の良さを阻害することなく、いかにグラフパターンの表現力を向上させるかという課題に取り組んだ。その結果、ブロック保存型グラフパターン(BP グラフパターン)と呼ばれるグラフパターンのクラスを提案し、パターン照合・発見アルゴリズムといったデータマイニングにおける最も基本的なアルゴリズムを開発した。BP グラフパターンの表現力は、BPO グラフパターンよりも真に大きい。したがって、実世界のデータに対して、より複雑なパターンを効率良く発見することが可能になった。

(4) 一般的なグラフ構造がどれだけ木構造に近いかを表すグラフ理論的な指標として、グラフの木幅がある。我々はグラフパターンの一般化を目的とし、グラフの木幅を考慮したグラフパターンを提案し、そのグラフパターン照合問題に対するアルゴリズムを提案した。実際、化学化合物のほとんど全てが小さい木幅のグラフで表現できることが知られており、本結果により化学化合物のほとんど全てを効率の良いグラフパターンマイニングの対象とすることができる。

(5) スケジューリング，地図情報や複雑なテーブルを表現するためのグラフパターンのクラスとして、区間グラフパターン，TTSP グラフパターン，平面図パターンを設計し、効率の良いパターン照合・発見アルゴリズムを提案した。

(6) パターン発見アルゴリズムの効果的な利用として、ダークネット(未使用の IP アドレス空間)で観測されたデータのスクリーニング手法を提案した。インターネットにおける脅威の一つとしてボットネットなどによる不正アクセスがある。その脅威を早期に検知するためには、ダークネットなどのネットワーク観測による脅威の傾向把握が必要である。本研究では、脅威の傾向を明確にするために、ありふれた不正パケット群を、頻出時系列パターン発見によって取り除くデータスクリーニング手法を提案し、実データに対して満足すべき効果が得られることを確認した(図 3)。

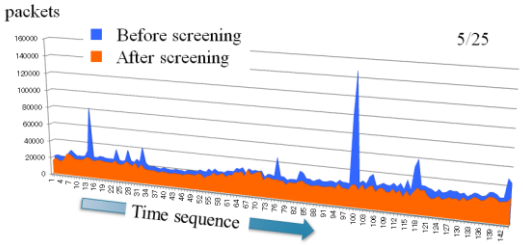

図 3 ダークネットにアクセスした不正パケットに対し、自動的に脅威の低いパケットを取り除く手法を提案した。青線と赤線はそれぞれ提案手法適用前後のパケット数を表す。

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１５件）

① Takashi Yamada and Takayoshi Shoudai, Efficient Pattern Matching on Graph Patterns of Bounded Treewidth, Electronic Notes in Discrete Mathematics, 査読有, to appear, 2011.

② Hitoshi Yamasaki, Takashi Yamada, and Takayoshi Shoudai, Graph Pattern Matching with Expressive Outerplanar Graph Patterns, Lecture Notes in Electrical Engineering, 査読有, 70, 2010, pp. 231-243.

③ Yuko Itokawa, Koichiro Katoh, Tomoyuki Uchida, and Takayoshi Shoudai. Algorithm Using Expanded LZ Compression Scheme for Compressing Tree Structured Data, Lecture Notes in Electrical Engineering, 査読有, 52, 2010, pp. 333-346.

④ Satoshi Kawamoto, Yusuke Suzuki, and Takayoshi Shoudai, Learning Characteristic Structured Patterns in

Rooted Planar Maps, Proc. International MultiConference of Engineers and Computer Scientists 2010 (IMECS2010), 査読有, Vol. I, 2010, pp. 465-470.

⑤ Satoshi Matsumoto, Yusuke Suzuki, Takayoshi Shoudai, Tetsuhiro Miyahara, and Tomoyuki Uchida, Learning of Finite Unions of Tree Patterns with Repeated Internal Structured Variables from Queries, IPSJ Transactions on Mathematical Modeling and its Applications, 査読有, 2, 2009, pp.127-137.

⑥ Hitoshi Yamasaki, Yosuke Sasaki, Takayoshi Shoudai, Tomoyuki Uchida, and Yusuke Suzuki, Learning block-preserving graph patterns and its application to data mining, Machine Learning, 査読有, 76, 2009, pp.137-173.

⑦ Ryoji Takami, Yusuke Suzuki, Tomoyuki Uchida, and Takayoshi Shoudai, Polynomial Time Inductive Inference of TTSP Graph Languages from Positive Data, IEICE TRANSACTIONS on Information and Systems, 査読有, E92-D, 2009, pp.181-190.

⑧ Hitoshi Yamasaki and Takayoshi Shoudai, A Polynomial Time Algorithm for Finding a Minimally Generalized Linear Interval Graph Pattern, IEICE TRANSACTIONS on Information and Systems, 査読有, E92-D, 2009, pp.120-129.

⑨ Hitoshi Yamasaki and Takayoshi Shoudai, Mining of Frequent Externally Extensible Outerplanar Graph Patterns, Proc. 7th International Conference on Machine Learning and Applications (ICMLA'08), 査読有, 2008, pp.871-876.

⑩ Satoshi Kawamoto, Hitoshi Yamasaki, and Takayoshi Shoudai, A linear time isomorphism algorithm for circular-arc graphs with connected domination number greater than 3, Proc. 11th JAPAN-KOREA Joint Workshop on Algorithms and Computation (WAAC 2008), 査読有, 2008, pp.123-130.

〔学会発表〕(計７件)

① Hisashi Tsuruta, Frequent Sequential Pattern Discovery for Data Screening, International MultiConference of Engineers and Computer Scientists 2011 (IMECS2011), 平成23年3月18日, ロイヤルガーデンホテル, 香港.

② 山田 貴志, 辺縮約制約を持つ部分 k-木への辺縮約問題に対する多項式時間アルゴリズム, 情報処理学会アルゴリズム研究会, 平成23年1月12日, 愛媛大学.

③ 鶴田 悠, ダークネット観測データの時系列パタン発見によるスクリーニングについて, 電子情報通信学会インターネットアーキテクチャ研究会, 平成22年6月17日, 京都大学学術情報メディアセンター.

④ Takashi Yamada, A Polynomial Time Algorithm for Finding a Minimally Generalized Externally Extensible Outerplanar Graph Pattern, The 6th Workshop on Learning with Logics and Logics for Learning (LLLL 2009), 平成21年7月6日, 京大会館, 京都府.

⑤ 川本 哲, 平面図パタンの多項式時間学習アルゴリズムに関する研究, 情報処理学会九州支部火の国情報シンポジウム, 平成21年3月14日, 九州産業大学情報工学部.

６．研究組織

(1)研究代表者

正代　隆義 (SHOUDAI TAKAYOSHI)
九州大学・システム情報科学研究院・准教授
研究者番号：５０２２６３０４

(2)研究分担者

内田　智之 (UCHIDA TOMOYUKI)
広島市立大学・情報科学研究科・准教授
研究者番号：７０２６４９３４

(3)連携研究者

鈴木　祐介 (SUZUKI YUSUKE)
広島市立大学・情報科学研究科・助教
研究者番号：１０３９８４６４

松本　哲志 (MATSUMOTO SATOSHI)
東海大学・理学部・准教授
研究者番号：３０３０７２３５