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　17 日現在

研究種目：若手研究(B)
研究期間：2009～2010
課題番号：20700143
研究課題名（和文）グラフの局所的構造に基づく大規模半構造データからの高速パターン発見

研究課題名（英文）Rapid pattern discovery from huge semi-structured data based on local structure in graph

研究代表者
坂本　比呂志（SAKAMOTO　HIROSHI）
九州工業大学・大学院情報工学研究院・准教授
研究者番号：50315123

研究成果の概要（和文）：
従来は極めて困難であった規模の半構造データから，特徴的なパターンを高速に発見するスケーラブルなマイニング技術を実現した．本手法は，この種の問題のボトルネックであった部分グラフ同型判定問題を回避しながら，高い精度でパターンの発見を行うことが可能である．

研究成果の概要（英文）：
For the difficult mining problem from huge semi-structured data, we developed a scalable method to find characteristic patterns immediately. By this method, we can obtain interesting patterns with high accuracy avoiding the subgraph isomorphism problem, which is a bottleneck of such problem.

交付決定額

（金額単位：円）

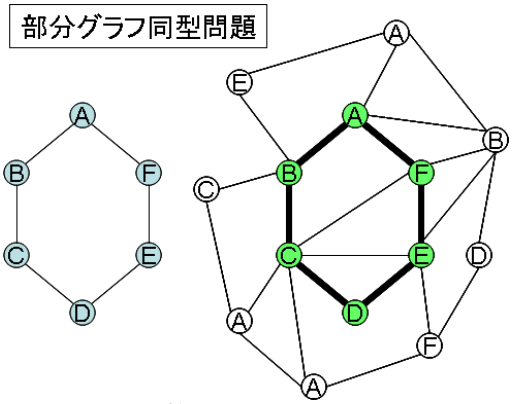
	直接経費	間接経費	合　計
2009 年度	1,000,000	300,000	1,300,000
2010 年度	900,000	270,000	1,170,000
年度			
年度			
年度			
総　計	1,900,000	570,000	2,470,000

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：ウェブマイニング，知識発見，半構造データ，機械学習

１．研究開始当初の背景

グラフ構造から特徴的なパターンを発見することは，XML データからのパターン発見をはじめ，遺伝子ネットワークからの構造抽出や化学物質の分子構造予測など応用例の枚挙にいとまがなく，日々蓄積され巨大化するこれらのデータ群から有用なパターンを高速かつ自動的に取り出す研究が盛んに行われている．

関連研究：**頻出共通部分木発見**では，木構造データから頻出部分木を高速発見するアルゴリズムが提案されているが，データが順序木に限定されている．より複雑な構造からのマイニングとして，一連の**グラフマイニング**による頻出部分構造発見では，一般のグラフから頻出部分グラフを発見する手法が提案されているが，パターンをデータから陽に探



し出すための計算量が増加し，大規模化が困難である．

着想に至った経緯：これらの関連研究をふまえて，**データのさらなる大規模化と発見の高速化を行うためには，NP-完全問題である部分グラフ同型問題を直接扱わず，近似的な解の発見を目指すべき**である．そして，データ群を高い精度で説明する近似パターンを高速に発見することができれば，正確なパターンの予測に十分に役立つと考えられる．

２．研究の目的

本研究が目的とするデータ群は，リンク構造を含んだ XML によって記述される有向グラフのクラスと同等である．このようなリッチなデータ構造からパターンを発見することは，グラフアルゴリズムにおける部分グラフ同型問題やその亜種である NP-完全問題に帰着されるため，その根本的解決はほぼ不可能である．この問題に対して，従来技術では，取り出すパターンの複雑さや対象データの規模を制限することで困難性を回避しようとしている．例えば，**ウェブコミュニティの発見**では，簡単に計算可能なグラフ中の辺密度が高い部分グラフをコミュニティとみなすことで，時間計算量を現実的な範囲に収めている．また，対象グラフを隣接行列で表現する**アプリオリ・ベースのグラフマイニング**[b]では，領域計算量が入力の二乗に比例してしまうため大規模化が困難である．パターン発見におけるこれらの困難性は，グラフに埋め込まれたパターンを完全な形で抽出するための計算コストに起因する．そこで本手法では，高速に計算可能なタグやパスの頻度，接点間の距離などの**グラフの局所的構造から学習を行うことでグラフの特徴量を抽出し，それらの局所構造同士の距離計算によって，求める特徴を緩やかに満たす近似的なパターンを生成**する．特に，これらの計算に必要な時間/領域-計算量を入力サイズの線形程度に抑えることが大規模化の実現に必要な目標である．そこで，本研究課題として次の研究項目 A,B,C を達成することを目指す．

(A) **畳み込みカーネルによる局所構造抽出**：カーネル法は，特徴空間上のベクトル同士の類似度を再帰的計算によって高速に抽出するためのテクニックである．畳み込みカーネルは，その値を単純なベクトルの内積計算によって求める手法であり，本研究では，これをグラフの特徴的な局所構造の抽出に応用する．局所構造を頻度などの基準で抽出した場合には，どの部分にも含まれるような自明な構造ばかりが抽出されてしまうおそれがあるが，特徴量によって分類することで，意味のある局所構造を優先して取り出すことがねらいである．

(B) **XML 索引による構造間の距離計算**：抽出した局所構造がどのように結合してグラフに埋め込まれているかを計算することは，部分グラフ同型問題と等価である．そこで，局所構造間の距離が高速に計算できれば，それらのグラフ上における大まかな配置を求めることが出来る．したがって，それらのうち互いに近いもの同士がパターンを形成していることが予測できる．本研究では，グラフ上の接続関係を判定するアルゴリズムを，接点間の距離が計算できるものへと改良してこの問題を解決する．

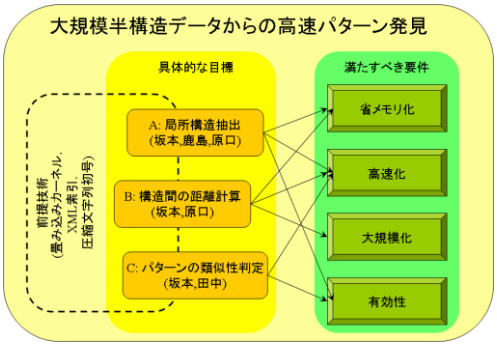
(C) **圧縮文字列照合によるパターンの類似性判定**： グラフ構造だけではなくテキストや属性値の類似性にも着目して特徴パターンの抽出精度を向上させる．同じ構造を持つパターン内の属性値などは，値は異なるもののその書式は似通っていると考えられる．このような仮定から，あらかじめ属性やテキスト部分を圧縮保存しておき，必要に応じてその類似性を計算する．

このようにして，関連技術を発展させて目的のパターンを抽出する．また，一部あるいは全体が似ているパターンがデータの別々の場所から抽出された場合，それらは類似性を判定することでさらに一般的なパターンに汎化できる．

３．研究の方法

本研究課題で提案するグラフの局所的構造に基づく大規模半構造データからの高速パターン発見は，研究目的の欄で述べたように A,B,C の 3 つの研究目標を骨子として持つ．そしてこれらの目標は，次世代の高速パターン発見が満たすべき要件である**「省メモリ化」，「高速化」，「大規模化」，「有効性」**の 4 項目を達成することからなっている．図にそれらの相関図を示す．本研究課題は，この研

究計画に従って遂行される．以下に各項目毎にこの研究計画の具体的進め方を示す．



４．研究成果

従来手法では取り扱いが極めて困難な規模の半構造データから，特徴的なパターンを高速に発見するスケーラブルなマイニング技術を実現した．この目的を達成するために，情報処理技術(畳み込みカーネル，XML 索引，データ圧縮アルゴリズム)を援用し，グラフ構造からのパターン発見のボトルネックである，部分グラフ同型判定を回避しつも高い精度でのパターン発見を行う手法を開発した．最終年度は，昨年度に引き続き以下の項目を達成するためのプログラムを完成した．

(A) **畳み込みカーネルによる局所構造抽出：**前年度に開発した省メモリ化を達成するための不要パターンをあらかじめ取り除く前処理手法をグラフ構造上で実装した．

(B) XML 索引による構造間の距離計算：データを分割することで，大規模 XML データに対する高速索引付けを可能とした．また，ノード間の距離計算手法を改良し局所構造間距離計算を実現した．これらの性能をデータ分割の手法をＰＣクラスタ上で実装し，その性能を確認した．

(C) **圧縮文字列照合によるパターンの類似性判定：**XML データは頻繁にデータの更新が起こるため，データの変更に対して影響が少ない圧縮法が望ましい．そこで，前年度までに開発した適応型圧縮アルゴリズムを大規模データに適用し，規模耐性が高いことを確認した．

以上のように，最新のプログラムをＰＣクラスタ等の分散システム上で実装し，本研究の成果を国際会議や論文誌，国内研究会等で公表した．今後は，家庭用ＰＣなどの環境で動作するより軽量なアルゴリズムを開発していく．

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　2 件）

① 丸山史郎, 坂本比呂志
データ圧縮による大規模情報検索の実現と関連情報マイニングへの応用―テキストの特徴をつかまえる圧縮技術―
情報管理, 53(5):233-240, 依頼原稿, 2010.
査読無。

② Yushi Nakamura, Toshihiko Horiike, Tetsuji Kuboyama, Hiroshi Sakamoto
Extracting Research Communities from Bibliographic Data
KES Journal, 査読有, to appear

〔学会発表〕（計　11 件）

① T. Kuboyama, K. Ito, K. Hirata, H. Sakamoto Predicting Mutation Trend of Influenza A Virus through Dimensionality Reduction in Hamming Metric on HA Amino Acid Sequences Annual International Conference on Bioinformatics and Computational Biology (psoter, CD-ROM), 2011

② ″Shirou Maruyama, Masahiro Baba, Hiroshi Sakamoto, Kunihiko Sadakane and Masafumi Yamashita″ ″A Practical Random Access to Grammar-Based Compression″ The 4th Annual Meeting of Asian Association for Algorithms and Computation, 2011.

③ ″Shirou Maruyama, Masaya Nakahara and Hiroshi Sakamoto″ ″An Online Algorithm for Lightweight Compression of Highly Repetitive Text ″ The 4th Annual Meeting of Asian Association for Algorithms and Computation, 2011

④ 丸山史郎，馬場雅大，岸上直也，坂本比呂志 文法型圧縮法の全二分木表現による符号化とランダムアクセス手法の提案, 第 134 回アルゴリズム研究会，2011

⑤ 中村優士，堀池寿彦，久保山哲二，坂本比呂志 関連語の自動選定による論文コミュニティ抽出技術の改良，第 81 回 SIG-FPAI 研究会，2011

⑥ 坂本比呂志，岸上直也，中原昌也， 丸山史郎 長い部分文字列を検索するための文法圧縮索引 2011 年冬の LA シンポジウム，2011.

⑦ 馬場雅大，丸山史郎，坂本比呂志， 定兼邦彦，山下雅史 文法圧縮に基づいた圧縮データの自己索引構造化の提案, 2011 年冬の LA シンポジウム，2011

⑧ 岸上直也，中原昌哉，丸山史郎，坂本比呂志 長いパターンを検出するための文法圧縮に基づく索引構造 第 133 回アルゴリズム研究会，2011

⑨ 坂本比呂志，検索可能な文法圧縮の実現―文字列間類似度の高速計算―″FIT2010 イベント企画 Science と Engineering をつなぐ『Art』を求めて－ERATO 湊離散構造処理系プロジェクトシンポジウムーポスターセッション″，2010.

⑩ 馬場雅大，丸山史郎，坂本比呂志，小野廣隆，定兼邦彦，山下雅史　　Edit Sensitive Parsing を用いた文法圧縮に基づく省スペースな索引構造 -理論編-　　人工知能学会第 78 回 SIG-FPAI 研究会，2010.

⑪ 丸山史郎，馬場雅大，坂本比呂志，小野廣隆，定兼邦彦，山下雅史　　Edit Sensitive Parsing を用いた文法圧縮に基づく省スペースな索引構造 -実験による評価-　　人工知能学会第 78 回 SIG-FPAI 研究会，2010

〔図書〕（計　0 件）
〔産業財産権〕
○出願状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等
http://www.donald.ai.kyutech.ac.jp/~hiroshi/

６．研究組織
研究代表者
坂本比呂志（SAKAMOTO　HIROSHI）
九州工業大学・大学院情報工学研究院・准教授・
研究者番号：50315123