科学研究費補助金（特別推進研究）公表用資料
〔事後評価用〕

平成１７年度採択分
平成２０年　３月３１日現在

研究課題名（和文）
知識基盤形成のための大規模半構造データからの超高速パターン発見
研究課題名（英文）
Efficient Pattern Discovery from Massive Semi-Structured Data for Knowledge Infrastructure Formation on the Web

研究代表者
有村 博紀（Hiroki Arimura）
北海道大学・大学院情報科学研究科・教授

研究の概要：
本研究では，World Wide Web（WWW，ウェブ）に代表されるネットワーク上の膨大な電子情報に内在する知識を発見することを目的として，超高速半構造パターン発見技術と，これと組み合わせて知識基盤形成を行うための周辺技術の研究開発を行った．成果として，理論的性能保障をもつ超高速半構造マイニングアルゴリズムや，知識索引技術や知識連携技術を研究開発し，ウェブ空間における知識基盤形成支援システムのアーキテクチャと実装技術を確立した．

研　究　分　野：**情報科学**
科研費の分科・細目：**情報学・知能情報学，メディア情報学・データベース**
キ　ー　ワ　ー　ド：ウェブ，半構造マイニング，超高速パターン発見，知識索引

１．研究開始当初の背景

現在 World Wide Web（WWW，ウェブ）は，人類が歴史上かつて体験したことがない膨大な知の集積となっている．しかし，現在の情報技術(IT)では，情報検索などの限られたアクセス手段しかないのが実情であり，ウェブ上に集積された膨大な知識を互いに関連付け，外在化して理解し，知識を取り出すための新しい情報アクセス技術の開発が緊急の課題となっている．一方，データ量の急速な増大を背景として，1990 年代初頭から，データベースに蓄積された大量のデータから，有用な規則性やパターンを半自動的に取り出す方法の研究であるデータマイニング(data mining)が登場した．

データマイニングは，現在，理論と応用の両面で活発な研究が行われているが，大規模半構造データは，(i)膨大な量の，(ii)多様な構造をもつ，(iii)非定型データの集積であり，従来のデータマイニング手法を，そのまま適用することはできない．そのため，来るべき知識集約的社会において，巨大なウェブから有用な知識を効率よく取り出すための技術の確立が急務となっている．

２．研究の目的

これに対して，本研究では大規模知識ネットワークを対象としたデータマイニング，すなわち大規模知識ネットワークマイニングについて研究する．本研究の目的は以下の 3 点にまとめられる．

(a) 膨大な半構造データから知識をパターンや規則としてとりだす超高速な半構造マイニングエンジン技術を開発する．特に，計算量に徹底的にこだわりながら，理論的な性能保障をもつ高速アルゴリズムを研究開発する．

(b) この半構造マイニング技術を，現実の多様な半構造データに適用するための周辺技術を開発し，ウェブ空間における知識基盤形成システムのアーキテクチャと，その実現のための実装技術を確立する．

(c) これらの技術をもとに半構造マイニングエンジンを実装し，知識基盤形成のための周辺技術とともに世界に公開する．さらに大規模半構造データからの知識基盤形成可能性に関して，具体的な領域を選び実証実験を行う．

３．研究の方法

大規模知識ネットワークマイニングの鍵になる技術として，申請者等がこれまでの研究で開発してきた超高速最適化半構造データマイニングの枠組みを採用し，これを多様な半構造データに拡張する．本研究では，基盤技術と周辺技術の両面から，次の研究項目に統一的に取り組む．

(A)基盤技術として，知識ネットワークに対する高速な半構造データマイニング・エンジンを開発する．また，理論的解析によってその本質的難しさを明らかにし，その一般理論を構築する．

(B)現実の多様な半構造データに適用するための周辺技術を開発する．　特に，知識断片を有機的に連携する知識連係機構と，発見したパターンを圧縮したまま格納し自由な操作を実現する大規模知識索引技術を開発する．

(C)さらに，これらを有機的に結合することにより，大規模知識基盤形成システムの実現方法を明らかにすることを目標とする．

本研究の特色として，第 1 に実際の重要性にもかかわらず，系統的に研究されていなかった大規模知識ネットワークマイニングにとりくむ．第 2 にコンテンツベースの新しいアクセス手法として，追求する．第 3 にデータマイニング・エンジンの開発だけでなく，知識基盤形成の 3 つの基本的問題すべてを統一的に扱う． 第 4 に計算時間に徹底的にこだわり，計算量理論とアルゴリズム設計技法を駆使して，きわめて高速なアルゴリズムを追求する．

4．研究の主な成果

本研究では，半構造マイニングの基盤技術として，湊の ZBDD に基づく大規模知識索引技術 VSOP，両者を結合した LCMoverZBDDs，ウイルス学における知識発見などの代表的な半構造データからの知識発見技術等の成果を生んだ．詳細は次に示す．

4.1 超高速パターン発見アルゴリズムの開発．申請者らが先に開発した最右拡張法（図1）を，系列や，木，グラフ等の半構造データの族に対して拡張し，多項式時間遅延計算量という良好な性能保障をもつ超高速パターン発見アルゴリズムを開発した(MaxMotif 法，MaxFlex 法，CloATT 法, MaxGeo 法)．これらは自明でない半構造データの族に対する世界初の性能保障付き極大パターン発見アルゴリズムである[1,3]．

4.2 知識基盤形成の周辺技術の開発．分担者の湊による ZBDD 技術に基づく大規模知識索引技術を開発した．大規模論理回路処理の基盤技術であった ZBDD をデータマイニングに拡張して大規模知識索引技術 VSOP を実現し，頻出アイテム集合や，対称パターン，包含パターンといった高度なパターンの発見を圧縮データ上で効率よく実行可能な技術を開発した[2,4]．

4.3 プロトタイプ構築と実領域での知識発見システムの実証実験．研究メンバーの連携により，ZBDD-growth（湊・有村）と LCMoverZBDDs（湊・宇野・有村），ウイルス遺伝子からの知識発見（伊藤・湊）[2]，半構造情報検索技術（喜田・有村）[5]等，ウェブ上の高速パターン発見と，大規模知識索引，生命科学への応用など知識発見の新しい可能性を示す研究成果を得た．

4.4 開発した技術の実装と公開．宇野・有村の高速極大パターン発見技術 LCM の高度化版と合わせて，湊の ZBDD に基づく大規模知識索引技術 VSOP，両者を結合した LCMoverZBDD などを，当初の計画に従い実装・公開した．後半では，情報発信や知識普及にも注力し，全体として，学術的成果の追求，成果の周知，社会への還元などの活動も行った．

5．得られた成果の世界・日本における位置づけとインパクト

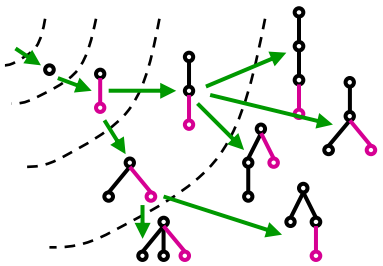

図1：最右拡張技法による半構造マイニング

現在，半構造マイニング技術は，米国や欧州の大学研究機関（イリノイ大学，フロンフォーファー研究所，パリ大学等）で盛んに研究されているが，多くの結果はアドホックな技術の提案にとどまる．これに対して，本研究の一連の結果は，計算量理論とアルゴリズム設計技法を駆使することで，理論的な性能評価をもち同時に実世界の大規模半構造データに適用可能な超高速半構造マイニング技術を開発した．とくに，主要な半構造データ族である系列や，木，グラフの族に対して，世界で初めて理論的性能保障をもつ極大パターン発見アルゴリズムを与えている．

本研究の進展により，基盤となる技術である最右拡張技術はウェブ上のオンライン引用 DB(Google scholar)で 220 件超の，ZBDD 技術は 260 件超の文献引用をうける等，本分野の基礎技術となっている．また，学会・産業界でも新しいタイプのウェブからの知識発見技術として注目されており，経済産業省「情報大航海」の基盤技術開発課題として，実装・公開されるなど，日本発の新たな情報基盤技術となっている．

6．主な発表論文

（研究代表者は太字、研究分担者には下線）

1. **H. Arimura**, T. Uno, An Efficient Polynomial Space and Polynomial Delay Algorithm for Enumeration of Maximal Motifs in a Sequence, *J. Combinatorial Optim.*, Vol.13, pp.243-262, 2007.
2. S. Minato, Kimihito Ito, Symmetric Item Set Mining Method Using Zero-suppressed BDDs and Application to Biological Data, *Trans. the Japanese Society for Artif. Intelligence,* 有, Vol.22, No.2, 156-164, 2007.
3. **H. Arimura**, Efficient Algorithms for Mining Frequent and Closed Patterns from Semi-structured Data,, *Proc. 12th Pacific-Asia Conf. on Knowledge Discovery and Data Mining*, 2008. (invited talk)
4. S. Minato, K. Satoh, T. Sato, Compiling Bayesian Networks by Symbolic Probability Calculation Based on Zero-suppressed BDDs, *Proc. IJCAI-2007*, pp.2550-2555, 2007.
5. S. Morinaga, **H. Arimura**, 他 3 名, Key Semantics Extraction by Dependency Tree Mining, *Proc. 11th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining* (KDD'05), pp.666-671, 2005.

ホームページ等

http://www-ikn.ist.hokudai.ac.jp/tokusui/