様式 C-7-2

自己評価報告書

平成　２３年　４月　２７日現在

機関番号：１０１０１
研究種目：基盤研究（Ａ）
研究期間：2008　～　2011
課題番号：20240014
研究課題名（和文）　大規模知識基盤形成のための次世代半構造マイニング技術の研究

研究課題名（英文）　Next-Generation Semi-structured Data Mining for Large-Scale Knowledge Base Formation

研究代表者
有村　博紀（ARIMURA HIROKI）
北海道大学・大学院情報科学研究科・教授
研究者番号：20222763

研究分野：情報科学
科研費の分科・細目：情報学・知能情報学
キーワード：知識発見とデータマイニング

１．研究計画の概要

本研究においては，ネットワーク上の大規模半構造データに内在する知識をパターンや規則としてとりだすことが可能な超高速な半構造マイニングエンジン技術を開発し，これ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する．さらに，開発した基盤技術と周辺技術の実装を行い，インターネット上の大規模半構造データからの知識発見実験を行うことを目的とする．

２．研究の進捗状況

（研究全体の進捗状況）平成20年度は，大規模知識基盤形成システムのための技術調査と，基盤技術の開発，環境整備を行った．平成21年度と平成22年度は，調査整備に基づき，大規模知識基盤形成システムのための基盤および応用技術の開発と，その理論解析，最適化を行った．具体的には，これまでの３年間の研究期間で，次の項目に関して研究・開発を行った．

（１）超高速半構造マイニングエンジンの研究開発（有村・宇野・喜田）．高速な重み付きマイニングや，２次元画像，幾何グラフ，対する最適化データマイニング技術を開発した．超高速半構造マイニングエンジンの研究開発として，菱形と，２部グラフ，多部グラフエピソードに対する深さ優先方式の頻出エピソードマイニング手法の開発と解析を行った．これらに関して，連携研究者が情報処理学会平成22年度IPSJ論文船井若手奨励賞(Katoh, Hirata, Arimrua, 2011.03.25)を受賞した．

（２）確率的情報処理スキーマと半構造データマイニングの結合の研究（有村・喜田・湊・宇野）．ストリームハードウェア上の柔軟な超高速半構造パターン照合方法（FPT2011）や，確率的刈込み接尾辞木を用いた系列予測手法(DS2008)を開発した．さらに，根付き木構造上の階層ベイズモデルを導入し，効率よい各種学習アルゴリズムを開発した．実証実験として感染症の流行解析やスパムフィルタリングの実証実験を行った(GIW2009)，連携研究者が情報処理学会 平成22年度情報処理学会山下記念研究賞を受賞した(柳橋，2009年SIGBIO研究会)．

（３）半構造データマイニングの一般理論の構築（有村・宇野・湊）．従来の半構造データマイニング手法の構造を一般化し，系列，木，グラフの部分族に対して，アクセス可能集合族に基づく極大部分構造マイニングアルゴリズムの一般的構成法を与えた．この結果は，本分野の一級の国際会議であるSIAM Data Mining国際会議で発表された(Arimura, Uno, SDM2009)．また，時系列マイニングに関しても，効率よくマイニング可能なエピソードのクラスの特徴付けとアルゴリズム構成法を与えた(Katoh, Hirata, PAKDD2008)．

（４）大規模知識基盤形成システムのための半自動知識連係技術の研究開発．ネットワーク上の膨大なウェブページからの情報抽出技術開発として，半構造ストリームに照合アルゴリズムのハードウェア上の超高速実装方法や，圧縮パターン照合技術の開発を行った．可変長固定長符号や接尾辞木を用いた高速検索が可能な圧縮方式(Kida, Yoshida IEEE DCC2009, 2010, 2011; SPIRE2010; Uemura, Arimura, CPM2011, to appear)や，正規表現や木構造パターンのための高速ビット並列検索方式(Kaneta, Minato, Arimura, SPIRE2010,

IWOCA2010)，そのFPGAやGPUなどの並列ハードウェアへの実装方法(FPT2010)などの成果を得た．（伊藤・喜田・湊・有村）
（５）ZBDDに基づく高速な大規模知識索引技術を開発した（湊・喜田・有村）．分担者の湊が開発したZBDDに基づく大規模知識索引技術に基づいて，データベース系列から，知識索引と頻出アイテム集合発見を密結合したLCMoverZDD技術（Minato, Arimura, PAKDD2008）や，与えられた生起パターン表現をもつ頻出アイテム集合発見手法の開発(Minato, Uno, SDM2010)，大規模な系列データのための大規模知識索引技術SeqBDDの理論的解析とシステム構築等を行った．さらに分担者の湊は，巡回の離散構造上へのBDDの拡張に成功した(SAT2011, to appear)．一連の研究に関し，2010年信学会 情報・システムソサイエティ論文賞(湊,有村,2010.06.01)を受賞した．
（６）研究開発と並行して，開発した知識発見技術のプロトタイプ実装を行った．さらに，理論解析と計算機実験による評価を行い，これらを元にさらなる最適化を行った．また，実装と計算機実験を行った．(平成21年度)
情報処理学会 平成22年度山下記念研究賞(柳橋，2010.05.22)等を受賞した．

３．現在までの達成度

① 当初の計画以上に進展している。
この４年間で，当初の予定通り，本研究の中核である超高速半構造マイニングアルゴリズムの開発と一般的設計手法の確立に成功しており，さらに，期待以上の結果として，発展形である時系列データ上でのグラフパターン発見や，知識索引との連携，大規模プロトタイプ実装等の成果が得られ，国際的な発表・出版を行うことができた．

４．今後の研究の推進方策

研究項目（１）～（６）の各項目について，当初の計画以上に進展しているので，そのまま研究計画を遂行する．特に，最終年度の2011年度は，(1)を中心に，(2)では木構造上の確率モデルの効率よい学習，(3)では列挙可能性と計算効率の一般的特徴づけ，(4)では圧縮検索技術と超高速ビット並列検索，(5)では各種半構造データへの知識索引の一般化，(6)プロトタイプの評価実験に重点を置いて研究開発を行う．

５． 代表的な研究成果
（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 23 件）

(1) T. Katoh, H. Arimura, An Efficient Depth-first Search Algorithm for Extracting Frequent Diamond Episodes from Event Sequences, IPSJ Online Trans., 有, Vol.3, 2010, 1-12.

(2) T. Uemura, D. Ikeda, T. Kida, and H. Arimura, Unsupervised Spam Detection by Document Probability Estimation with Maximal Overlap Method, Trans. of the Japanese Society for Artificial Intelligence, 有, Vol. 26, 2010, 297-306.

(3) Y. Kaneta, S. Minato, H. Arimura (他2名), Dynamic Reconfigurable Bit-Parallel Architecture for Large-Scale Regular Expression Matching, Proc. the 2010 IEEE Int'l Conference on Field-Programmable Technology , 有, IEEE, 2010, 21-28.

(4) S. Minato, T. Uno, Frequentness-Transition Queries for Distinctive Pattern Mining from Time-Segmented Databases, Proc. SIAM Int'l Conference on Data Mining, SDM 2010, 有, SIAM, 2010, 339-349.

(5) Y. Kaneta, S. Minato, and H. Arimura, Fast Bit-Parallel Matching for Network and Regular Expressions, Proc. the 17th Symp. on String Processing and Information Retrieval (SPIRE2010), 有, LNCS, Vol.6393, Springer, 2010, 372-384.

(6) S. Yoshida and T. Kida, An Efficient Algorithm for Almost Instantaneous VF Code Using Multiplexed Parse Tree, Proc. of Data Compression Conf. 2010, 有, IEEE, 2010, 219-228.

(7) H. Arimura and T. Uno, Polynomial-Delay and Polynomial-Space Algo- rithms for Mining Closed Sequences, Graphs, and Pictures in Accessible Set Systems, Proc. 2009 SIAM Int'l Conf. on Data Mining 2002 (SDM'09), 有, 2009, SIAM, 1087-1098.

〔学会発表〕（計 38 件）

(1) 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 (PAKDD2008), LNCS, Vol.5012, Springer, 2008.（招待講演）

〔図書〕（計 0 件）
〔その他〕