様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　１２　日現在

機関番号：　１４３０１
研究種目：　基盤研究(B)（一般）
研究期間：　2014～2016
課題番号：　２６２８００８５

研究課題名（和文）順序関係が成立する属性を持つデータからの閉集合を用いた知識発見

研究課題名（英文）Knowledge Discovery Methods based on Closed Set Construction for Data with Attributes Whose Values are from Ordered Sets

研究代表者

山本　章博（Yamamoto, Akihiro）

京都大学・情報学研究科・教授

研究者番号：３０２３０５３５

交付決定額（研究期間全体）：（直接経費）　　12,200,000円

研究成果の概要（和文）：本研究の目標は，属性が備える順序関係と閉集合を利用した知識発見手法を開発することであった．閉集合による知識発見は２項関係データの双クラスタリングとよばれる手法の一種であるので，行列の因子分解と探索による閉集合構成，様々な双クラスタリング手法で得られた閉集合間の関係解析を行った．また，木構造データの部分構造間の半順序を用いた木構造の主成分分析手法，自然言語シソーラスにおける語彙の順序関係を用いたシソーラス拡張，整数計画法による木構造の共通部分の高速発見手法という成果を得た．

研究成果の概要（英文）：The goal of this research is to develop methods for knowledge discovery that uses both orders for attribute values and closed sets of from mixed data. Since knowledge discovery by closed set is one kind of technique called bi-clustering of binary relational data, we obtained a method for bi-clustering for matrix factorization and search of matrices. Also we analyzed relationship between closed sets obtained by various bi-clustering method. Moreover, we developed methods for principal component analysis of the tree structure using the partial order of the substructure of tree data, the thesaurus expansion using the order relation of the vocabulary in the natural language thesaurus, and discovery methods of the similarity of the tree structure data with the integer programming.

研究分野：　知能情報学

キーワード：　データマイニング　２項関係　閉集合　順序関係　木構造　主成分分析

１．研究開始当初の背景

World Wide Web などの広い意味でのデータベース内に蓄積されている大量かつ多様なデータから有用な知識を効率的に導出する知識発見手法が重要になっている．特に WWW を通じて蓄積されるデータは，データの作成・取得時刻という数値データが自然と定義でき，さらに自然言語文や Web ページ間のリンク構造など離散値属性が容易に付加できるという特長を持つ．つまり，数値デ属性と離散値属性を組にした数値・離散値混合データ（以下，「混合データ」とよぶ）は容易に蓄積される．混合データの典型例がネットワークのアクセス記録データであり，アクセス時刻とプロトコルの組からなっている．このようなデータから有用な知識発見を行うには，数値データの特性と離散値データの特性を利用することが不可欠である．特に，時刻属性を利用した知識を抽出する方法が重要となる．

数値データに対して時刻属性を利用した知識抽出は，時系列解析の分野ですでに長い研究の蓄積がある一方，離散データや混合データに対する時系列解析は，データ空間の構造解析の手法が数値解析ほど強力ではないため，研究が進んでいないのが現状である．

２．研究の目的

本研究の最終的な目標は，混合データから属性が備える順序関係と閉集合の両方を利用した知識発見手法を開発することにある．順序関係の例としては，時刻を表現する数値データの全順序，木構造・グラフ構造データの属性として利用する部分構造（パターン）間の半順序，自然言語データの属性としての語彙のシソーラスを対象とする．研究代表者が従前に進めてきた研究で得られた閉集合を用いた知識発見に関する成果を発展させ，順序関係の利用手法を融合させることにより，時系列データの転換点発見，自然言語文書から新語挿入によるシソーラスの更新を含む新たな知識発見手法を開発する．

３．研究の方法

次のように設定する基本原理に基づいて，課題(1)～(4)を研究する．

[基本原理] データのある属性の領域が順序空間であると仮定した上で，データ空間から構成された閉集合を順序空間上に射影し，射影された閉集合間に順序関係を定義した上で，不要な閉集合の像を除き，有用な閉集合の像を抽出する．

課題(1) 順序関係を持つ属性を用いた形式概念解析順序空間を領域とする属性が一つの場合を扱う．具体的には，属性が数値である場合について，「小数点以下 n 桁目までの数字列」がブール値属性になり，木構造データに対してはデータである木に含まれる部分木（パターン）を属性が属性となる．そして，このように拡張した属性から構成した閉集合を順序空間に射影した場合に，新たに閉集合間に成立してしまう順序，元の閉集合間で成立していた順序で外さなければならないものを整理した上で，どのような閉集合の像と像間の順序関係が成立すべきかを明らかにする．

課題(2)構造データ間の順序関係の深化：構造を持つ離散データ間の順序関係について，前研究を展開する．最も単純な木構造ですら，多様な順序を定義可能であることがわかっている．情報が欠落している木構造データを扱えるように順序関係を拡大することを目指す．

課題(3) 閉集合の射影の選択アルゴリズムの実装：課題(1)で得られた理論を基盤にして満たすように[基本原理]を実装する．初期解は，従来研究で知られている高速閉集合アルゴリズムを利用して閉集合をすべて発生させ，課題(1)で得られた理論を用いて，閉集合の順序空間への射影の取捨選択を行う，というものである．最初に最
も単純な場合として，意味的に順序を表すカテゴリカル属性だけを含む形で実装し，次に時刻などの数値データを用いた転換点発見，単語集合データ，木構造データとそれらの順序関係を扱える形に拡大することで，課題(4)での実データの適用が可能な形にする．次に，高速閉集合アルゴリズムに閉集合の射影の選択を埋め込む方法を考察する．

課題(4) 実データへの適用のよる実用性検討：順序関係を持つ属性を表現する適切なデータ構造を検討した上で，閉集合の時刻属性への射影を求めることにより，どのような知識が得られるかどうかを検証する．その結果を検討することでアルゴリズムと仕様に対して改良すべき点を検討する．

４．研究成果

研究目的に述べた課題(1)と(3)については，同期させながら研究を進めた．まず具体的な対象とするデータを選定し，そこから構成される閉集合の順序関係を考察することから開始した．データとしては，公共事業と入札企業の関係を表すデータを用いた．このデータには入札時刻という順序属性が考えられ，閉集合間の順序関係は入札履歴を可視化するために有効である．この着想に基づいて，当初は[基本原理]に忠実に実装していたが，課題(3)で述べたように閉集合間の順序関係には閉集合の枝刈りが必要であることから，閉集合の構成，順序化，枝刈りを同時に実装する手法について，行列の因子分解と探索による閉集合構成を利用することで考案した．

また[基本原理]を再検討した結果，それが主成分分析(PCA)に類似していることに気付き，木構造間の編集距離を用いた PCA 方式を提案した．まず，ユークリッド空間内のデータに対する PCA を抽象化することで，公理的な主成分分析を定義した．そして，木構造デ

ータ空間内の”直線”として部分木を採用し，個々の木データと”直線”としての部分木の距離を編集距離で測ることとした．

また，研究進捗が進展により，閉集合による知識発見が２項関係データの双クラスタリングとよばれる手法の一種であることが判明した．そこでより一般的な問題である双クラスタリングで得られた閉集合間の関係の解析を進めることとした．形式概念解析が２項関係データを対象とすることから直感的にはデータは２次元空間内の点で表現され，閉集合とは矩形である．当初計画では複数の閉集合を順序関係を持つ方の属性に射影した上で，その影の関係を分析することあったが，閉集合の２次元空間での位置関係を分析することとした．位置関係の中には，射影した場合の関係も含まれるので，当初計画よりより一般的なものになる．解析の方法としては，数理論理学や統計理論を範にして，真の閉集合間の関係を一つ定めて人工的にデータを発生させ，閉集合列挙アルゴリズムによって真の関係が復元されるかどうかを検討する，という方法を採用した．結果として，この分析結果がアルゴリズムの設計者が想定する真の関係を明らかにする，ということがわかった．この結果は課題(3)に対する部分的な解にもなってる．

課題(2)については，木構造間の編集距離が成立するか否かを判定するための手順を整数計画法を用いて高速に計算する手法を提案した．入力となる２つの木を構成する接点と辺の間の関係を，整数変数とそれらを用いた不等式で表現して，整数計画法のソルバに入力可能な形に変換する．編集距離は木構造間の順序関係をもとに定義され，その順序関係のバリエーションを不等式のバリエーションで表現するため，この方法は木構造間の順序関係解析に応用可能と考えている．さらに，当初考案した手順に動的計画法を組み合わせることによって計算が劇的に改善されることを実証した．

課題(4)については，実データに対する実用性の検討については，閉集合を用いて自然言語文データからシソーラスを拡張する方法について，実験を繰り返した上で学術論文としてまとめた．

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）

〔雑誌論文〕（計　3 件）

1．Seiichi Kondo, Keisuke Otaki, Madori Ikeda, Akihiro Yamamoto, Fast Computation of The Tree Edit Distance Between Unordered Trees Using IP Solvers, Lecture Notes in Computer Science,査読有，8777, 156-157, Springer, 2014.
https://link.springer.com/chapter/10.1007/978-3-319-11812-3_14
2．Tomoya Yamazaki, Akihiro Yamamoto, Tetsuji Kuboyama, Tree PCA for Extracting Dominant Substructures from Labeled Rooted Trees, Lecture Notes in Computer Science，査読有，9356，316-323，Springer，2015.
https://link.springer.com/chapter/10.1007/978-3-319-24282-8_27
3．Madori Ikeda and Akihiro Yamamoto, Extending Various Thesauri by Finding Synonym Sets from a Formal Concept Lattice, 自然言語処理，査読有,24,(印刷中) 2017.

〔学会発表〕（計 10 件）

1．Issei Hamada, Kouichi Hirata, Tetsuji Kuboyama, Takaharu Shimada, Agreement -Subtree Mapping Kernel and Leaf-Path Kernel for Phylogenetic Tree Reconstructed from Nucleotide Sequences, Workshop on Graph-based Algorithms for Big Data and its Applications (GABA2014), 2014.
2．Yuma Ishizaka, Takuya Yoshino, Kouichi Hirata, Anchor Alignment Problem for Rooted Labeled Trees, Workshop on Graph-based Algorithms for Big Data and its Applications (GABA2014), 2014..
3．Shoichi Nishimura, Keisuke Otaki, Madori Ikeda, Ryo Yoshinaka, Akihiro Yamamoto, Takeaki Uno, Updating a Closed Itemset Family Based on Inclusion Relations, Workshop on Graph-based Algorithms for Big Data and its Applications (GABA2014), 2014.
4．狩山和亮，Marco，Cuturi，山本章博，久保山哲二，福元健太郎，密度優先探索に基づくコミュニティ抽出と入札データ分析への応用，人工知能学会 第 97 回人工知能基本問題研究会(SIG-FPAI)，2015.
5．山崎朋哉，山本章博，久保山哲二，木構造データからの主成分抽出，2015 年度人工知能学会全国大会．
6．山崎朋哉，山本章博，久保山哲二，Tree PCA による任意形状の木構造を抽出するアルゴリズム，第 99 回人工知能基本問題研究会，2016.
7．西村翔一，吉仲亮，山本章博，閾値の変化に対する高速なグラフ研磨の再計算手法，第 99 回人工知能基本問題研究会，2016.
8．芳野拓也，平田耕一，Tai マッピングの根無し木への拡張，第 99 回人工知能基本問題研究会，2016.
9．山浦智佳子，小林靖明，山本章博，久保山哲二，クラスタ構造を仮定した場合の双クラスタリングアルゴリズムの解析，第 103 回人工知能基本問題研究会(SIG-FPAI)，2017.
10．Hong Eunpyeong，小林 靖明，山本 章博，整数計画法による木間距離の計算を高速化するための新しい定式化，第 103 回人工知能

基本問題研究会(SIG-FPAI), 2017.

〔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

６．研究組織
(1)研究代表者
山本 章博（YAMAMOTO, Akihiro）
京都大学・大学院情報学研究科・教授
研究者番号：３２０３０５３５

(2)研究分担者
伊藤 公人（ITO, Kimihito）
北海道大学・人獣共通感染症リサーチセンター・教授
研究者番号：６０３９６３１４

平田　耕一（HIRATA, Koichi）
九州工業大学・情報工学研究院・教授
研究者番号：２０２７４５５８

久保山　哲二（KUBOYAMA, Tetsuji）
学習院大学・計算機センター・教授
研究者番号：８０３０２６６０

吉仲 亮（YOSHINAKA, Ryo）
京都大学・大学院情報学研究科・助教
(2016.3 まで)
東北大学・大学院情報科学研究科・准教授
(2016.4 から)
研究者番号：３２０３０５３５

(3)連携研究者
AVIS, David
京都大学・大学院情報学研究科・元特定教授
(4)研究協力者
小林靖明（KOBAYASHI Yasuaki）
京都大学・大学院情報学研究科・助教
(2016.9 から)
池田真士里（IKEDA, Madori）
京都大学・大学院情報学研究科・博士後期課程
(2016.3 まで)
大滝啓介（OTAKI, Keisuke）
京都大学・大学院情報学研究科・博士後期課程
(2016.3 まで)
狩山和亮（KARIYAMA, Kazuaki）
京都大学・大学院情報学研究科・修士課程
(2015.3 まで)
山崎朋哉（YAMAZAKI, Tomoya）
京都大学・大学院情報学研究科・修士課程
(2016.3 まで)
西村翔一（NISHIMURA, Shoichi）
京都大学・大学院情報学研究科・修士課程
(2016.3 まで)
山浦智佳子（NISHIMURA, Shoichi）
京都大学・大学院情報学研究科・修士課程
(2017.3 まで)