様式 C-19

科学研究費補助金研究成果報告書

平成２２年　６月　４日現在

研究種目：基盤研究（B）
研究期間：2007～2009
課題番号：19300046
研究課題名（和文）　頻度に基づく下方集合の高速探索を利用した構造データからの知識発見

研究課題名（英文）　Knowledge Discovery from Structured Data with Efficient Methods of Searching Lower Sets Based on Frequency

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

研究成果の概要（和文）：本研究は，データベース内に蓄積されている様々なデータ構造を採る大量のデータの中に潜む有用な知識を効率的に発見する手法の基盤構築を目標とした．データのなす束構造の様々な性質を明確にした上で，束の下方集合の高速探索を利用した機械学習・知識発見アルゴリズムを設計した上で，一部は理論上の未解決問題の解決や自然科学への適用を試みた．特に下方集合が持つ閉集合としての性質が，基礎理論構築と応用の両側面で有用であることが明確になった．

研究成果の概要（英文）：This research project aimed at constructing foundations of efficient knowledge discovery from data of large scale and of various data types, stored in databases. We clarified various properties of the lattice structure consisting of various types of structured data, designed machine learning / knowledge discovery algorithms by using efficient search of lower sets in the lattices, and applied some of them to theoretical open problems or to problems in science other than informatics. We found that the property of lower sets as being closed sets is useful for both fundamental theories and applications.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	3,800,000	1,140,000	4,940,000
2008 年度	3,000,000	900,000	3,900,000
2009 年度	3,200,000	960,000	4,160,000
年度			
年度			
総　計	10,000,000	3,000,000	13,000,000

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：学習と発見，機械学習，知識発見，帰納論理，計算代数

１．研究開始当初の背景

データベース内に蓄積されている大量のデータに潜む有用な知識を効率的に発見する手法(知識発見手法)の重要性は，データの量のみならず多様性が増す今日では益々増している．関係データベースからの頻度に基づく知識の発見手法は“大量データからの知識発見”という概念そのものを確立し，現在に至るまで知識発見研究の基本手法の一つとして用いられている．その後，10 年経って，

単調命題論理式を用いた計算論的学習理論を用いた理論が与えられた．これらは，当時の主流データ管理方法である関係データベースを対象とし，関係スキーマの属性を命題変数とみなすことを基本としている．

一方で，学術研究や社会活動の急激なデジタル化に伴って,XML/HTML 文書をはじめ,自然言語の構文解析データ，生命科学に関するデータや化学式に関するデータなど構造を持ったデータ(構造化データ)が増加し，蓄積したデータベースの整備が進められている．それらの中には部分的に数値データを含むこともある．このような現状では，関係データベースからの知識発見手法だけでは限界があり，構造化データを対象とした知識発見の基礎理論を構築する必要がある．

２．研究の目的

従来研究では関係データベースを対象としていた頻度に基づく知識発見について，本研究では構造化データを対象とするための基礎理論の構築を目標とする．具体的には，次の問題(1)(2)を解決する．

(1) 従来整備されてきた関係データベースからの知識発見の理論では，発見される知識の表現言語が命題変数の連言である．発見の対象が構造化データのデータベースの場合には，抽出される知識もデータ構造を表現できるような複雑な「式」を表現可能な知識表現言語を用いる必要がある．

(2) 知識表現の複雑化にあわせて，データから発見される知識の正当性の保証や，発見という行為自体の正当性の保証，知識発見手法の効率化なども再構成する．

３．研究の方法

問題(1)，(2)に対して，計算論的学習理論を基盤とし，下方集合の探索とその高速化を用いて解決する．下方集合とは，与えられ得る「式」全体が束構造をなすときに，ある式よりも詳細な意味を持つ「式」が全て含まれている部分束である．このとき，下方集合を定義する束構造に頻度の概念を導入し，開発する知識発見手法に対して，発見される知識の正当性，発見行為自体の正当性，知識発見の効率化の正当性を与える理論を構築する．

４．研究成果

知識表現手法となる「式」として，一階述語論理の項，時系列事象間の関係を表す有向グラフ（エピソード)，一般の有向グラフ，木，楽譜を表す MusicXML，イデアルを表す多項式，形式言語の生成文法を対象とした上で以下の研究を行った．

(1) 下方集合を利用した機械学習・知識発見のための基礎的・理論的な性質の分析を多方面から行った．特徴的な成果として以下を得ている．

①木構造データの機械学習の基礎となる木構造データ間の距離に関する性質を網羅的に明らかにした．

②木構造データをあらかじめ圧縮しておくことで機械学習を高速化するための一般性をもった手法を考案し，実装した．

③以下の(3)で述べる文脈自由文法を用いて設計したカーネル関数を一般化し，演繹的な導出を用いて構造データから機械学習するためのカーネル関数の形に定式化した．

③統計的推論に対して機械学習の研究で一般的な VC 次元の分析を行った．

(2) 木構造，エピソード，楽譜を表す MusicXML，グラフを対象とした知識発見のための理論的かつ網羅的研究を行った上で，以下の応用を試みた．

①楽譜を表す MusicXML データからの知識発見を用いてクラシック音楽の作曲家の特徴を抽出を試みた．

②グラフからの知識発見を用いてインフルエンザデータから地域間伝播を抽出することを試みた．

(3) 形式言語の生成文法の機械学習を応用する研究を行い，以下の特徴的な成果を得た．

①機械学習手法としてサポートベクトルマシンを選び，その一般的性質の分析を行った．さらに，サポートベクトルマシンについて，文脈自由文法における導出を「式」とみなすことにより，カーネル関数を設計を行い，実問題として RNA 配列データの分類問題に適用することでその有効性を示すことに成功した．

②形式言語の性データからの機械学習手法を固体物理の結晶解析に応用可能な形に展開した．

(4) 多項式イデアルの一般形である閉集合を機械学習，知識発見に応用し，以下の成果を得た．

①有界でない個数の閉集合の和集合を学習するアルゴリズムを理論的に構成し，機械学習の理論研究で未解決の問題を解決した．

②閉集合をデータマイニングに応用することにより，公開ソフトウェアとその開発メーリングリストの相互関係からソフトウェア構造の抽出に成功した．

③分散配置した関係データベースから，情報遺漏をできるだけ少なくするデータマイニングアルゴリズムを，閉集合を用いて設計した．

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 17 件）

① Sankoh, H., Doi, K., and Yamamoto A.:

An Intentional Kernel Function for RNA Classification, Lecture Notes in Artificial Intelligence, 査読有, 4755, 2007, 281-285

② Takashi Katoh, Kouichi Hirata and Masateru Harao: Mining Frequent Diamond Episodes from Event Sequences, :Lecture Notes in Artificial Intelligence, 査読有, 4617, 2007, 477-488

③ Takashi Katoh, Kouichi Hirata, Masateru Harao, Shigeki Yokoyama and Kimiko Matsuoka: Extraction of Sectorial Episodes Representing Changes for Drug Resistance and Replacements of Bacteria, Proc. IEEE/ICME International Conference on Complex Medical Engineering, 査読有, 2007, 304-309

④ 土井晃一郎, 山下哲矢, 田中孝侑, 山本章博:内包カーネル関数, 人工知能学会論文誌, 査読有, 23, 2008, 185-192.

⑤ Itsuo Takamatsu, Masanori Kobayashi, Hiroo Tokunaga and Akihiro Yamamoto: Computing Characteristic Sets of Bounded Unions of Polynomial Ideals, Lecture Notes in Artificial Intelligence, 4914, 2008, 318-32

⑥ Tetsuji Kuboyama, Kouichi Hirata and Kiyoko F. Kinoshita:An Efficient Unordered Tree Kernel and Its Application to Glycan Classification,Lecture Notes in Artificial Intelligence, 査読有, 5012, 2008, 184-195

⑦ Takashi Katoh and Kouichi Hirata: A Simple Characterization on Serially Constructible Episodes, Lecture Notes in Artificial Intelligence, 査読有,5012, 2008, 600-607

⑧ Murakami, S., Doi, K., and Yamamoto, A.: Finding Frequent Patterns from Compressed Tree-structured Data, Lecture Notes in Artificial Intelligence, 査読有,5255, 2008, 284-295

⑨ Yuuichi Kameda, Hiroo Tokunaga and Akihiro Yamamoto: Learning bounded unions of Noetherian closed set systems via characteristic sets, Lecture Notes in Artificial Intelligence, 査読有, 5278, 2008, 98-110

⑩ Taku Aratsu, Kouichi Hirata, and Tetsuji Kuboyama: Approximating Tree Edit Distance through String Edit Distance for Binary Tree Code, Lecture Notes in Artificial Intelligence, 査読有, 5404, 2009, 93-104

⑪ Taishin Daigo and Kouichi Hirata: On Generating Maximal Acyclic Subhypergraphs with Polynomial Delay, Lecture Notes in Artificial Intelligence, 査読有, 5404, 2009, 181-192

⑫ Yohji Akama: Commutative Regular Shuffle Closed Languages, Noetherian Property, and Learning Theory, Lecture Notes in Computer Science, 査読有, 5457, 2009, 93-104

⑬ Takashi Katoh, Hiroki Arimura, and Kouichi Hirata: A Polynomial-Delay Polynomial -Space Algorithm for Extracting Frequent Diamond Episodes from Event Sequences, Lecture Notes in Artificial Intelligence, 査読有, 5476, 2009, 172-183

⑭ Kazuya Sata, Kouichi Hirata, Kimihito Ito, and Tetsuji Kuboyama: Discovering Networks for Global Propagation of Influenza A (H3N2) Viruses by Clustering, Lecture Notes in Artificial Intelligence, 査読有, 5712, 2009, 490-497

⑮ Takashi Katoh, Hiroki Arimura,and Kouichi Hirata: Mining Frequent Bipartite Episode from Event Sequences, Lecture Notes in Artificial Intelligence, 査読有, 5808, 2009, 136-151

⑯ Dinh Anh Nguyen, Koichiro Doi, and Akihiro Yamamoto: Discovering the Structures of Open Source Programs from Their Developer Mailing Lists, Lecture Notes in Artificial Intelligence, 査読有, 5808, 2009, 227-241

⑰ デ・ブレクト マシュー, 徳永浩雄, 山本章博: 代数学・数学基礎論における機械学習, 人工知能学会誌, 査読無, 24, 2009, 788-796

〔学会発表〕(計 32 件)

① Yohji Akama, Taufik Sutanto: Multiclass Multisurface Proximity Support Vector Machines, Fiftth Workshop on Learning with Logics and Logics for Learning (LLLL2007), 2007 年 6 月 18 日, 宮崎

② Takashi Katoh, Kouichi Hirata: Mining Frequent Elliptic Episodes from Event Sequences, Fiftth Workshop on Learning with Logics and Logics for Learning (LLLL2007), 2007 年 6 月 19 日, 宮崎

③ Takamatsu, I., Kobayashi, M., Tokunaga, H., and Yamamoto, A.: Computing Characteristic Sets of Bounded Unions of Polynomial Ideals, Fiftth Workshop on Learning with Logics and Logics for Learning (LLLL2007), 2007 年 6 月 19 日, 宮崎

④ 三功浩嗣, 土井晃一郎, 山本章博: RNA 配列を識別するための内包カーネル関数の設計, 人工知能学会 第 66 回人工知能基本問題研究会, 2007 年 7 月 14 日, 湯布院

⑤ 土井晃一郎, 山本章博: 内包カーネルとその応用, 人工知能学会 第 4 回人工知能学会データマイニングと統計数理研究会, 2007 年 7 月 26 日, 旭川

⑥ 荒津拓, 平田耕一: 単純で高速な木の類似性尺度, 人工知能学会 第 68 回人工知能基本問題研究会, 2008 年 1 月 17 日, 札幌

⑦ 大梧諦真, 平田耕一: 極大非巡回部分超グラフの列挙アルゴリズム」, 人工知能学会 第 68 回人工知能基本問題研究会, 2008 年 1 月 17 日, 札幌

⑧ 三功浩嗣, 土井晃一郎, 山本章博: RNA 識別における内包カーネルの性質, 人工知能学会 第 68 回人工知能基本問題研究会, 2008 年 1 月 17 日, 札幌

⑨ 徳永浩雄, 山本章博: 完備特徴例集合を用いた言語の有界和の正データからの学習 --多項式イデアルから木パターン言語へ--, 人工知能学会 第 68 回人工知能基本問題研究会 2008 年 1 月 16 日, 札幌

⑩ 赤間陽二, 上野康隆: 主成分分析の汎化誤差などについて, 人工知能学会 第 68 回 人工知能基本問題研究会, 2008 年 1 月 17 日, 札幌

⑪ 澤田石翔太, 三功浩嗣, 土井晃一郎, 山本章博: 内包カーネルと配列分割法を用いた RNA 識別, 情報処理学会 第 12 回バイオ情報学研究会, 2008 年 3 月 4 日, 福岡

⑫ Arnoldo Jose Muller Molina, Kouichi Hirata and Takeshi Shinohara: A Tree Distance Function Based on Multi-sets, the First International Workshop on Algorithms for Large-Scale Information Processing in Knowledge Discovery (ALSIP2008), 2008 年 5 月 20 日, 大阪

⑬ Taku Aratsu, Kouichi Hirata, and Tetsuji Kuboyama: Sibling Distance for Rooted Labeled Trees, the First International Workshop on Algorithms for Large-Scale Information Processing in Knowledge Discovery (ALSIP2008), 2008 年 5 月 20 日, 大阪

⑭ Doi, K. and Yamamoto, A.: Kernel Functions Based on Derivation, the First International Workshop on Algorithms for Large-Scale Information Processing in Knowledge Discovery (ALSIP2008), 2008 年 5 月 20 日, 大阪

⑮ 赤間陽二, 入江慶, 河村彰利, 上野康隆: 統計推論の VC 次元, 人工知能学会 第 70 回 人工知能基本問題研究会, 2008 年 7 月 4 日, 京都

⑯ 亀田勇一, 徳永浩雄, 山本章博: Noether 閉集合族の有限和の学習, 人工知能学会 第 70 回 人工知能基本問題研究会, 2008 年 7 月 4 日, 京都

⑰ 久野慎弥, 土井晃一郎, 山本章博:

MusicXML データからの頻出フレーズパターンの発見, 人工知能学会 第 70 回人工知能基本問題研究会, 2008 年 7 月 4 日, 京都

⑱ 大梧諦真, 平田耕一: 極大非巡回部分超グラフの辞書順列挙, 人工知能学会 第 70 回 人工知能基本問題研究会, 2008 年 7 月 4 日, 京都

⑲ 赤間陽二, 入江慶, 河村彰利, 上野康隆: 主成分分析のＶＣ次元, 日本応用数理学会 2008 年度年会, 2008 年 9 月 19 日, 柏

⑳ 木村 誠, 土井 晃一郎, 山本 章博: 構造を持つ楽曲データを対象とした質問学習にもとづく楽曲生成, 情報処理学会 第 77 回音楽情報科学研究会, 2008 年 9 月 22 日, 京田辺

21 Nguyen, V. A., Doi, K., Yamamoto, A.: Mining Maximal Tree Patterns with Subtree Constraint, The Third International Workshop on Data-Mining and Statistical Science (DMSS2008), 2008 年 9 月 25 日, 東京

22 グェン・ディン・アン, 土井晃一郎, 山本章博:メーリングリストに基づいた共同開発ソフトウェアの構造抽出, 人工知能学会 第 72 回 人工知能基本問題研究会, 2009 年 3 月 14 日, 東京

23 Taku Aratsu, Kouichi Hirata, Tetsuji Kuboyama:Local Frequency Distances for Rooted Ordered Trees, Sixth Workshop on Learning with Logics and Logics for Learning, 2009 年 7 月 6 日, 京都

24 Yuichi Kameda and Hiroo Tokunaga: Inferability of Unbounded Unions of Certain Closed Set Systems, Sixth Workshop on Learning with Logics and Logics for Learning, 2009 年 7 月 6 日, 京都

25 Takashi Katoh, Hiroki Arimura, Koichi Hirata: Mining Frequent k-Partite Episodes from Event Sequences, Sixth Workshop on Learning with Logics and Logics for Learning, 2009 年 7 月 6 日, 京都

26 Seishi Ouchi and Akihiro Yamamoto: Learning from Positive Data based on the MINL Strategy with Refinement Operators, Sixth Workshop on Learning with Logics and Logics for Learning, 2009 年 7 月 6 日, 京都

27 Yohji Akama: Gaussian Mixture Models and VC-Doimensions, The Fourth International Workshop on Data-Mining and Statistical Science, 2009 年 7 月 8 日, 京都

28 赤間陽二: 可換シャッフルクローズド正規言語, ネーター性, および極限学習, 人工知能学会第 74 回 人工知能基本問題研究会, 2009 年 9 月 14 日,広島

29 上野康隆, 赤間陽二: VC 理論と Wishart 行列の固有値の和の集中不等式, 第 12 回情報論的学習理論ワークショップ, 2009 年 10 月 19 日, 福岡

30 河東孝, 有村博紀, 平田耕一: 極小出現を用いた頻出多部エピソードの効率のよい発見アルゴリズム, 人工知能学会第 76 回 人工知能基本問題研究会, 2010 年 1 月 27 日, 熊本

31 林田崇佑, 柴田智博, 平田耕一: 局所ラベル木の文字列表現と編集距離, 人工知能学会第 75 回 人工知能基本問題研究会, 2010 年 1 月 27 日, 熊本

32 久野慎弥, 土井晃一郎, 山本章博: 分散データベースからの頻出飽和アイテム集合の発見, 人工知能学会第 76 回 人工知能基本問題研究会, 2010 年 3 月 17 日, 札幌

６．研究組織

(1)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

(2)研究分担者

平田 耕一 (HIRATA KOUICHI)

九州工業大学・情報工学研究院・准教授

研究者番号：20274558

土井晃一郎 (DOI KOICHIRO)

京都大学・情報学研究科・助教

研究者番号：10345126

徳永 浩雄（TOKUNAGA HIROO）
首都大学東京・理工学研究科・教授
研究者番号：30211395
(H21 より連携研究者)
赤間陽二（AKAMA　YOHJI）
東北大学・理学系研究科・准教授
研究者番号：30272454
(H22 より連携研究者)