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　３０　年　　６　月　　５　日現在

機関番号：　１２６１２
研究種目：　若手研究(B)
研究期間：　2014～2017
課題番号：　２６８７００１１

研究課題名（和文）決定グラフを用いた超大規模ヒッティング集合の列挙・索引化とその知識発見への応用

研究課題名（英文）A Large-scale Enumeration of Minimal Hitting Sets And Its Application to Knowledge Discovery

研究代表者

戸田　貴久（Toda, Takahisa）

電気通信大学・大学院情報理工学研究科・助教

研究者番号：５０４５１１５９

交付決定額（研究期間全体）：（直接経費）　　2,800,000円

研究成果の概要（和文）：本研究課題では、極小ヒッティング集合の列挙問題やその拡張問題（論理関数の双対化およびAll Solutions SAT問題）に対する実用上効率的な計算法の研究に取り組んだ。今回新たに考案した計算手法だけでなく、関連する主要な従来手法についても実装したソフトウェアを開発し、その有効性や特性を明らかにするために大規模な性能評価実験を実施した。それらの結果は誰でも利用できる形で公開している。また、知識発見への応用に向けて、本研究課題で得られた最新のソフトウェアをアイテムセットマイニング問題に適用し、実データを用いた評価実験を行い、今後の研究課題や各種の知見を得た。

研究成果の概要（英文）：This research project developed practically efficient methods for the following important problems: the minimal hitting set enumeration problem and extended problems such as the dualization of Boolean functions and the all solutions SAT problem.Not only novel methods proposed in this research, but also most major methods proposed before were implemented as public software, and their efficiency and characteristics were confirmed through comprehensive experiments. These results and software are publicly available on our website.
Toward applications to knowledge discovery, our software was applied to itemset mining problems based on the so-called declarative approach in data mining, and experiments using actual data taken from transaction databases were conducted, by which promising research problems in this approach and useful knowledge were obtained.

研究分野：　アルゴリズム

キーワード：　ヒッティング集合　列挙　アルゴリズム　二分決定グラフ　SAT　アイテムセットマイニング　双対化

１．研究開始当初の背景

ヒッティング集合は与えられた集合と交わりを持つ集合であり、**極小ヒッティング集合の列挙問題**は、包含関係に関して極小なヒッティング集合をすべて列挙する問題である（図１）。

図１．極小ヒッティング集合の列挙

この問題は、論理、データマイニング、人工知能などさまざまな分野にまたがる基礎的な問題であり、多くの重要な問題と等価であることが知られている。

理論的な観点からは計算量の解析が大きな注目を集めてきた。一方、極小ヒッティング集合を列挙することで、さまざまな計算問題を解くことができるので、実用的な計算手法の開発も重要である。この方面の研究は近年大きな盛り上がりをみせている。さまざまな手法が提案されているが、まだ十分ではなく、より効率的な手法の開発が求められている。

申請者は、極小ヒッティング集合の列挙問題に対して、**二分決定グラフ**と呼ばれるデータ構造を活用した新しいアルゴリズムを開発した。これまでの手法と大きく異なるアプローチであり、過去の研究の性能評価で用いられた多くのデータに対して著しい高速化がみられた。これにより、極小ヒッティング集合の列挙問題に対して二分決定グラフを用いる有効性が確認できた。

２．研究の目的

申請者はこれまでの研究で得られた成果を手掛かりにして、申請者の提案する計算アプローチのさらなる改善・発展を行い、極小ヒッティング集合の列挙問題に対する基盤的な計算技法の確立を目指す。また、極小ヒッティング集合の列挙問題の拡張にも取り組み、これまでの枠組みでは扱えない場合にも対応可能な、より一般的な問題に対する効率的な計算技法を開発する。そして、開発したプログラム、データ、各種の知見などを誰でも利用できる形で公開する。

３．研究の方法

(1) 提案手法のさらなる性能向上について

現状の計算法を構成する個々の要素技術の改善に取り組み、全体としての性能向上を目指す。具体的には、提案法の並列化、適切な変数順の算出法の考案、さらなる実験的・理論的解析を行う。

(2) 極小ヒッティング集合列挙の拡張

より複雑なデータにも対応させるため、問題を拡張する。その拡張問題に対して二分決定グラフを用いた計算法を与え、その性能の評価を行う。

(3) 知識発見などへの応用

知識発見分野などへの応用について検討を行う。良い応用問題が見つかれば、その問題に対して本手法を適用できるか考察する。その手法の実装を与え、実データを用いた性能評価実験を行い、有効性を確認する。また、関連する従来手法の調査も行う。

４．研究成果

(1) 提案手法のさらなる性能向上について

まず元の提案法の並列化に取り組んだ。元の提案手法において二分決定グラフは中心的な役割を果たしている。それは以下の２つの点からなる：大規模なデータを二分決定グラフとしてコンパクトに表現すること、そして二分決定グラフ上の演算を通して効率的なデータ処理を行うことである。ここで、二分決定グラフ上の演算とは、例えば、集合和、集合積、差集合、補集合の計算などのような集合演算を含むが、本研究では特に集合の**単項演算**（すなわち、１つの集合を入力として受け取り、１つの集合を出力する計算）に焦点を当て、（元の提案手法の内部処理で用いられた特定の単項演算だけでなく）さまざまな単項演算に対して適用可能な並列化の一般的枠組みを与えた。

さらに、並列化に頼らない高速化手法（すなわち、シングルスレッドのもとでの高速化手法）の開発にも取り組んだ。元の提案法のうち、もっとも計算コストのかかる部分は、入力データからすべてのヒッティング集合を表現する二分決定グラフ（BDD）を構築する部分である。従来は BDD 上の集合演算を繰り返し適用することで、この BDD 構築を実現していたが、この手法では最終的に得られる BDD だけでなく、途中で多数の BDD が生成される欠点があった。そこで、中間の BDD を経由しないで直接最終の BDD を構築する手法の開発に取り組んだ。この手法の基本的なアイディアは、**SAT ソルバー**と呼ばれる高速ソフトウェアを動作させながら同時に BDD を構築するところにある。この手法は、この後に取り組んだ極小ヒッティング集合の列挙問題の拡張問題に対して有用な手法であることが分かった（詳細は(4)において説明する）。

(2) 極小ヒッティング集合列挙の拡張

極小ヒッティング集合の列挙問題は、論理関数の観点からは、**単調論理関数の双対化問題**とみなすことができる。そこで、論理関数の観点から問題の拡張に取り組んだ。具体的には、単調とは限らない一般の論理関数を対象にした双対化（**論理関数の双対化**）に取り組んだ。

論理関数の双対化は、論理式の表現形式の変換（**CNF-DNF 変換**）の一種である。昔から計算量の解析など理論的な観点からよく研究されている重要な問題である。しかし、実

用上の効率性に重点を置いた最近の研究はほぼ皆無であった。そこで本研究では、(極小ヒッティング集合の列挙問題に対して私が提案した) 二分決定グラフに基づく計算手法を拡張することで、論理関数の双対化を計算する手法を提案した。

(3) 知識発見などへの応用

アイテムセットマイニングへの応用に取り組んだ。**アイテムセットマイニング問題**は、例えば顧客の商品購入履歴のようなデータベースが与えられるとき、一定の条件を満たす商品の組み合わせ (**アイテムセット**) を列挙する問題である。そのようなアイテムセットの典型例としては、頻繁に購入される商品の組み合わせに対応する**頻出アイテムセット**である。

アイテムセットに課す条件に応じて様々なアイテムセットマイニング問題が存在する。この分野の主要な研究方法は、それぞれのアイテムセットマイニング問題に対して、課される条件に特化した専用アルゴリズムを開発することであった。この特化型のアプローチは効率性の上で有効であるが、新しい条件や複数の条件の組み合わせなどを考えるとき、新たなアルゴリズム (そしてその実装) を開発する必要があり、開発コストが高くつく欠点がある。

この欠点に着目し解決を図る別アプローチが、近年、欧州のいくつかのグループにより活発に研究されている。その基本的なアイディアは、アイテムセットに課される条件を論理制約 (例えば命題論理式など) として記述し、その結果を汎用ソルバーで計算することである。

私は、この汎用ソルバーとして本研究の成果として得られた最新のソフトウェア (この後で説明する AllSAT ソルバー) を用いる研究に取り組み、実データを用いた基礎的な性能評価実験を行った。それにより、このアプローチを実用化する上で解決されるべき今後の課題が得られた。

例えば、従来の方法では、符号化された論理式のサイズが大きくなりすぎること、そしてそのように符号化された論理式から解を列挙することは時間がかかりすぎることなどである。そこでこの課題に対して、アイテムセットの基本的な条件を**疑似ブール制約**として直接表現できることに着目して、命題論理式まで変換するのではなく、疑似ブール制約から直接解を列挙するアプローチについて現在研究を進めている。

(4) 新たな展開：AllSAT

本研究課題の申請時には想定していなかった新しい展開について以下で説明する。

本研究課題の遂行過程において、ヒッティング集合の列挙問題を最も抽象化した形式の計算問題として **All Solutions SAT** (**AllSAT**) があることが分かった。この問題は、直観的には、「論理制約が与えられるとき、その解をすべて求める」ことを意味する。極小ヒッティング集合の列挙問題はそのようなタイプの問題の例であり、その他にも多くの列挙型の問題をカバーする。したがって、その応用範囲はとても広い。

例えば、アイテムセットマイニングへの応用やネットワークの検証問題にも適用可能である。これらの応用はそれぞれの研究分野 (上の例では、マイニング分野とネットワーク分野) において知られているのみであったので、本研究において ALLSAT の関連研究を網羅的に調査して、異なる分野における ALLSAT の応用事例をまとめた。

さらに、本研究では AllSAT 問題を効率的に解く手法を開発した。ここで私が土台にしたアプローチは、現代的な SAT ソルバーが基礎にする探索アルゴリズム **CDCL** と二分決定グラフ BDD を融合させた方法である。この枠組みは従来研究で提案されたものであったが、ALLSAT 問題の解法として一般に認知されておらず、他の従来手法との性能評価もなされていなかったので、その性能は未知であった。さらに、この計算の枠組みの単純な実装では実用上の性能が極めて低いことが事前実験で分かっていたので、この問題を解決するための高速化技法を開発した。また、メモリ使用量に上限も持たせる機能なども提案した。

この最新手法の性能を評価するために、主要な従来手法をすべて実装して大規模な比較実験を実施した。これまでは、従来手法の実装が公開されていなかったり、他の手法同士の大規模な比較実験もなされていなかったり、実験で使われるデータセットも限定的であったりしたので、今回の研究によってはじめて、それぞれの手法の特徴や実用上の効率性が明らかになった。特に、SAT ソルバーと二分決定グラフを融合させた手法は、ランダムインスタンスには弱いが、何らかの構造を持つインスタンスには、他の手法と比べて格段に良い性能を持つことが分かった。本研究で実装したプログラムや利用したデータセットはすべて、誰でも利用できる形で公開している。

本研究成果は、人工知能学会 第 103 回人工知能基本問題研究会(SIG-FPAI)において招待講演として発表した。

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 6 件)

① Takahisa Toda, Takeru Inoue, Exploiting Functional Dependencies of Variables in All-Solutions SAT Solvers, Journal of Information Processing, 査読有, Vol. 25 (2017), pp.459-468. http://doi.org/10.2197/ipsjjip.25.45

9

② Takahisa Toda, Takehide Soh, Implementing Efficient All Solutions SAT Solvers, ACM Journal of Experimental Algorithmics, 査読有, Vol. 21, No. 1, Article 1.12, 2016. http://dx.doi.org/10.1145/2975585

③ Takahisa Toda, Dualization of Boolean Functions Using Ternary Decision Diagrams, Annals of Mathematics and Artificial Intelligence, 査読有, Vol. 79, Issue 1, pp.229-244, 2017. http://dx.doi.org/10.1007/s10472-016-9520-z

④ Takahisa Toda, Koji Tsuda, BDD Construction for All Solutions SAT and Efficient Caching Mechanism, in Proc. of the 30th Annual ACM Symposium on Applied Computing (Track on Constraint Solving and Programming and Knowledge Representation and Reasoning), 査読有, pp. 1880-1886, April, 2015, Salamanca, Spain. http://dx.doi.org/10.1145/2695664.2695941

⑤ Takahisa Toda, Shogo Takeuchi, Koji Tsuda, Shin-ichi Minato, Superset Generation on Decision Diagrams, in Proc. of the 9th International Workshop on Algorithms and Computation (WALCOM 2015), LNCS, 査読有, Vol. 8973, pp. 317-322, Feb., 2015, Dhaka, Bangladesh. http://dx.doi.org/10.1007/978-3-319-15612-5_28

⑥ Shogo Takeuchi, Takahisa Toda, Shin-ichi Minato, A General Framework for Parallel Unary Operations on ZDDs, in Proc. of the fourth International Workshop on Algorithms for Large-Scale Information Processing in Knowledge Discovery, in conjunction with PAKDD 2014, LNCS, 査読有, Vol. 8643, pp.494-503, May, 2014, Tainan, Taiwan. http://dx.doi.org/10.1007/978-3-319-13186-3_44

〔学会発表〕(計 17 件)

① 戸田貴久, モデル検査における反例発見から反例列挙への拡張, 基盤(S) 離散構造処理系プロジェクト「2017 年度 初夏のワークショップ」, 北海道札幌市, (2017 年 6 月).

② 戸田貴久, (招待講演) 命題論理式を充足する変数割当の網羅的探索手法について, 人工知能学会 第 103 回人工知能基本問題研究会(SIG-FPAI), 大分県由布市, (2017 年 3 月).

③ 戸田貴久, 宋剛秀, 効率的な AllSAT ソルバーの実装と評価, 第 19 回プログラミングおよびプログラミング言語ワークショップ (PPL2017), 山梨県笛吹市, (2017 年 3 月).

④ 戸田貴久, 擬似ブール制約を解く, 情報系 WINTER FESTA Episode2, 東京 (2016 年 12 月).

⑤ 戸田貴久, 擬似ブール制約を解く, 基盤(S) 離散構造処理系プロジェクト「2016 年度 秋のワークショップ」, 北海道札幌市, (2016 年 11 月).

⑥ 戸田貴久, 井上武, AllSAT における変数従属関係, 基盤(S) 離散構造処理系プロジェクト「2016 年度 初夏のワークショップ」, 北海道札幌市, (2016 年 6 月).

⑦ 戸田貴久, 井上武, 変数間の支配関係に基づく論理式の全解列挙手法, 第 30 回人工知能学会全国大会, 1D5-OS-02b-6in2, 福岡県北九州市, 2016 年 6 月.

⑧ 戸田貴久, 宋剛秀, 全解列挙型 SAT ソルバー, 情報系 WINTER FESTA, 東京 (2015 年 12 月).

⑨ 戸田貴久, 宋剛秀, AllSAT ソルバの最近の進展, ERATO 湊離散構造処理系プロジェクト 2015 年度秋のワークショップ, 北海道千歳市, (2015 年 11 月).

⑩ 戸田貴久, 津田 宏治, BDD に基づく ALLSAT ソルバーを用いたアイテムセットマイニング, 第 29 回人工知能学会全国大会, 2H4-OS-03a-3in, 北海道函館市, 2015 年 5 月.

⑪ 戸田貴久, 津田宏治, ALLSAT のための BDD 構築および効率的なキャッシング技法, 人工知能学会 第 97 回人工知能基本問題研究会(SIG-FPAI), 大分県別府市 (2015 年 3 月).

⑫ Takahisa Toda, Koji Tsuda, "All Solutions SAT, BDD Compilation and Pattern Mining", JST ERATO Kawarabayashi Large Graph Project / Minato Discrete Structure Manipulation System Project Joint Workshop, Tokyo, Japan, Jan. 23, 2015.

⑬ Takahisa Toda, Koji Tsuda, Efficient Caching Mechanism in BDD Compilation, ERATO-ALSIP Special Seminar 2014, Kyoto, Japan, Dec. 13, 2014.

⑭ Takahisa Toda, Dualization Using Decision Diagrams and Its Application for Itemset Mining, Decision Diagrams in Optimization I, INFORMS2014, San Francisco, USA, Nov. 10, 2014.

⑮ 戸田貴久, DPLL 型 BDD 構築法の改善, ERATO 湊離散構造処理系プロジェクト 2014 年度秋のワークショップ, 北海道礼文島 (2014 年 9 月).

⑯ 戸田貴久，SAT ソルバーを用いた BDD 構築法，第 5 回 CSPSAT2 研究会，兵庫県神戸市（2014 年 8 月）．

⑰ 戸田貴久，メモリ階層を考慮した BDD 演算とその並列化，ERATO 湊離散構造処理系プロジェクト 2014 年度春のワークショップ，2014 年 4 月

〔図書〕（計 0 件）

〔産業財産権〕

○出願状況（計 1 件）

名称：状態遷移評価装置、状態遷移評価方法及び状態遷移評価プログラム
発明者：戸田　貴久、井上　武
権利者：同上
種類：特許
番号：特願 2016-101190
出願年月日：２０１６年５月２０日
国内外の別： 国内

○取得状況（計 0 件）

〔その他〕
ホームページ等
http://www.sd.is.uec.ac.jp/toda/index-ja.html

６．研究組織
(1)研究代表者
戸田　貴久　（TODA, Takahisa）
電気通信大学・大学院情報理工学研究科・助教
研究者番号：50451159