様式 C-19

科学研究費補助金研究成果報告書

平成 23 年 6 月 10 日現在

機関番号：62615
研究種目：若手研究(B)
研究期間：2007～2010
課題番号：19700017
研究課題名（和文）　効率的な極大極小元列挙アルゴリズムのための新しい理論構築とその実用化
研究課題名（英文）　A theoretical Approach to Efficient Maximal/Minimal Enumeration Algorithm and its Applications
研究代表者：　宇野 毅明（UNO TAKEAKI）
国立情報学研究所・情報学プリンシプル研究系・准教授
研究者番号：00302977

研究成果の概要（和文）：
極大(極小)解とは、他の解に含まれない（含まない）解であり、多数の解の中で性質の良いものたちの集合である。極大・極小解は一般に近接しておらず、効率的な探索は難しい。本研究では、逆探索や疎性の利用といったアルゴリズム手法を用いて効率的手法の構築法を研究し、多目的最適化、完全列、密部分グラフ、あいまい頻出集合、ディスタンスヒエディタリグラフといった構造に対する効率的な列挙アルゴリズムを与えた。

研究成果の概要（英文）：
Maximal (minimal) solutions are those not included in (not including) other solutions. Generally speaking, enumeration of maximal (minimal) solutions is not easy since they usually have no neighboring relations. In this research, we research efficient schemes of enumeration based on algorithmic technologies such as reverse search and the use of sparsity. As a result, we developed efficient enumeration algorithms for multi criteria optimization problem, perfect sequence, dense subgraph, ambiguous frequent itemset, and distance hereditary graphs.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	900,000	0	900,000
2008 年度	800,000	240,000	1040,000
2009 年度	800,000	240,000	1040,000
2010 年度	800,000	240,000	1040,000
年度			
総　計	3,300,000	720,000	4,020,000

研究分野：総合領域
科研費の分科・細目：情報学・情報学基礎
キーワード：列挙、アルゴリズム、極大、極小、双対化

１．研究開始当初の背景

双対化問題とは，与えられた集合族の，全ての要素と交わりを持つ極小な集合を全て見つける問題である．双対化問題は，グラフ理論でのハイパーグラフ双対化，組合せ最適化での極小集合被覆，論理での極小ヒッティングセットなどと等価な問題であり，他分野にまたがる非常に基礎的な問題である．また応用も多く，学習理論，データマイニング，最適化，述語論理などの分野に応用を持つ．双対化は，単調な集合族(ある集合が集合族に含まれるとき，その部分集合も全て含まれるような集合族)の，極大元の族を入力して極小元の族を求める問題，連言標準形で与えられ

た論理式を選言標準形に変換する問題などとも等価であるため，現実の問題でも，双対化を有効に使うことで効率的に解ける問題が多い．

双対化問題は計算量理論の面からも興味深い問題である．入力と出力の大きさ(N とする)の多項式時間で終了するアルゴリズムが存在するかどうかがいまだ未解決である．その一方で，Kachyan らにより N の log N 乗時間で終了するアルゴリズムが発見されており，このようなアルゴリズムが知られていない NP 完全問題とは異なる特徴がある．現在，時間・空間計算量で最適なアルゴリズムは，明治大学の玉木氏により提案されたものであり，また，多項式時間で解けるサブクラスの研究が東京大学の牧野氏らによって行われている．計算量にかかわる分野では，日本が世界をリードしていると言えるだろう．しかし，上記のアルゴリズムは，実際の計算にはあまり有効ではないことが，計算実験により確認されている．大きさが数十の問題であっても非常に時間がかかり，大きな問題はとても現実的な時間内に解けない．ヒューリスティックな手法を用いたアルゴリズムの研究もあるが，多くの解候補の探索や極小性の判定に時間がかかるため，中規模の問題を，時間をかけて解くことはできるが，巨大な実データは解けない．発見した解をメモリに蓄える必要があるため，多大なメモリを使用するという弱点もある．双対化はサブルーチンとして何度も呼び出されることが多く，中規模の問題を極短時間で解く必要もあるのだが，既存の手法はこのような目的に対してもさほど有効ではない．

２．研究の目的

本研究課題では，このような巨大な問題を短時間で，中規模問題を極短時間で解く，メモリ消費量の少ないアルゴリズムを構築し，効率良い実装を開発することを目的とする．メモリ消費量の少ない双対化アルゴリズムとして，発見した解をメモリに蓄えない深さ優先的なアルゴリズムが知られている．しかし，発見した解が極小解であるかどうかの判定に時間がかかるため，高速ではない．申請者は過去の研究で，極小性の判定を高速に行う技術を開発しており，この技術を大規模なデータにも対応できるよう改良することで，大きな問題を短時間で解くアルゴリズムの開発を行う．

３．研究の方法

現実の巨大データは疎であることが多い．このため，疎なデータに対して効率良く動くよう，疎性を利用したアルゴリズムの設計を行うと，巨大なデータに対する計算時間を劇的に減少できる．また，通常の列挙アルゴリズムは再帰型の計算を行い，レベルの増加に伴い反復数が指数的に増えるため，深いレベルでの反復の計算時間が減少するよう，浅いレベルの反復で前処理を行うことで，劇的に計算時間を減少できる．申請者は，列挙アルゴリズムの高速化を多く研究しており，特に近年はデータマイニングなどでの大規模列挙問題を効率良く解く手法について研究を行ってきた．そのため，列挙アルゴリズムの高速化，巨大データの取り扱いに関しては，理論面からの高速化技術に対して知見がある．特に，頻出集合列挙というデータマイニングの最も基礎的な問題においては，2004 年に行われた国際的なプログラムコンテストで優勝した実績を持つ．

さらにこれらの技術を，双対化問題を一般化した問題に対しても適用する．双対化は集合束上の極大元の集合から，どの極大元にも含まれない極小元を列挙する問題として考えることができる．応用分野では，集合束以外の束に対して同様の問題が考えられているが，現実的に有効なアルゴリズムの研究は少なく，特に大規模なデータを扱えるような技術の開発は行われていない．本研究課題では，この種の問題に対しても，大規模データの取り扱いと効率の良い列挙法の技術を適用する．

４．研究成果

多目的最適化問題とは、大きくしたい評価値を複数持つような問題であり、与えられた制約条件を満たす中で、なるべく多くの評価値を大きくする問題である。今回は、解が n 次元ベクトルであり、評価値が線形、制約条件も線形である問題を考えた。解ベクトルの中で、制約条件を満たすような微細な変動をどのように加えても、必ずいずれかの評価値が悪くなってしまうようなものを極大解と定義する。制約条件を満たす解の集合は多面体を形成するが、極大解はその多面体の端点に対応する。この問題は古くから知られ研究が行われてきたが、完全に列挙する効率良い方法に関しては研究が少ない。今年度の研究では極大解の列挙問題に対して初めて出力数線形時間の逆探索アルゴリズムを開発した。これは、この問題に対する、世界で初めての多項式時間アルゴリズムである。また同時に、シンプルな解法では NP 困難問題に突き当たるため、多項式時間アルゴリズムは望めないことも合わせて示した。

頻出集合とは、各項目がアイテムの集合であるデータベースの多くの項目に現れるアイテム集合である。アイテム集合の族は単調生を満たすため、極大解が自然に定義され、その列挙も盛んに研究されてきた。今回は現実問題の応用から、包含関係にあいまいさを導入し、頻出集合を拡張した疑似頻出集合の

概念を導入した。曖昧性を許容するパターンマイニングに対して、この成果が初の、多項式性の意味で効率的なアルゴリズムとなっている。疑似頻出集合も単調生を満たすため、極大解が自然に定義でき、既存の手法が直接的に利用できることを示した。また、ある程度の大きさの疑似頻出集合を直接的に見つけるアルゴリズムも提案した。

また、ディスタンスヒエディタリーグラフという距離保存性を持つグラフクラスに対して、唯一的なコードを与えて同型生からくるゆらぎを排除する手法を考案し、同時にコードを列挙することでディスタンスヒエディタリーグラフを多項式時間で列挙するアルゴリズムを開発した。これは、頻出パターン発見における同グラフの利用を可能とし、同時に極大解の列挙も可能にした。

コーダルグラフとは、長さ４以上のサイクルが必ずショートカットを持つグラフのことを言う。本研究では、コーダルグラフに含まれる部分コーダルグラフの列挙問題に対して、指定した枝を含み、他の指定された枝を含まないようなもののみを列挙するアルゴリズムについて研究を行ない、遅延が多項式時間であるアルゴリズムを開発した。また、同時に同問題の数え上げが困難であることも証明した。コーダルグラフから極大なクリークを逐次的に取り除いて得られるクリークの列を完全列という。完全列はクリークを除去する列の中で極大なものである。本研究では、完全列をグラフ的に特徴付けることにより効率的な列挙手法を開発し、遅延が0(1)である、つまり計算量的には最適であるアルゴリズムを開発した。

データマイニングの問題で、巨大な文字列から類似する部分文字列の組を全て見つけ出す問題に対して、効率的なアルゴリズムを開発した。また、連続的ハミング距離というあたらしい距離を導入し、この距離のもとでは極大な類似文字列が単純な形で定義され、かつ前述のアルゴリズムに、速度を大きく損なうことない改良を加えることで列挙できることを示した。これにより、従来は不可能であった巨大なゲノムの類似性を導くことが

ハイパーグラフ双対化という、極大要素列挙の基礎的な問題に対して、高速な計算手法を開発した。既存の手法とは異なり大きくメモリ使用量が減少させることに成功しており、また計算速度についても、既存の実装を1000倍以上上回る高速化を実現している。これについては実装と実験が終了し、現在国際会議に投稿中である。

極大な要素を列挙するアルゴリズムは、ある種のオラクルを仮定することで一般的な枠組みで議論されることが多い。しかし、これらオラクルは巨大データの計算で行われる工夫と相性が悪いことが多く、ゆえに巨大データでの一般的な極大解列挙アルゴリズムはあまり発展してきていない。今年度の研究で、大規模計算に対応したオラクルを用いることで、大幅な速度向上を実現する方法を開発し、それを用いて極大要素を列挙するアルゴリズムの開発を行い、実装を終了した。

また、列挙だけでなく、極大元の数え上げについても研究を行った。その始まりとして、マッチングの数え上げに関して研究を行った。マッチングの数え上げは、一般的に非常に難しい問題と見なされているが、マッチングの数え上げに関しても、コーダルグラフなどの基本的なグラフクラスでも難しいことを、集合被覆の数え上げという困難性が示されている問題を帰着することで証明した。また、チェイングラフという非常に基礎的なクラスでは解けることを示した。また、この成果を拡張して、距離保存グラフやトレマイックグラフなどのクリーク幅が小さいグラフクラスにおいては、極大マッチングの数え上げが多項式時間で効率良く行えることを示した。この結果を拡張することで、それらのグラフでのパスやパスマッチングの数を数えることが多項式時間でできることを証明した。

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 15 件）

① Yoshio Okamoto, Takeaki Uno, “A polynomial-time-delay and polynomial-space algorithm for enumeration problems in multi-criteria optimization”, European Journal of Operational Research 210, pp. 48-56, 2011（査読有り）

②Takeaki Uno, “An Efficient Algorithm for Solving Pseudo Clique Enumeration Problem”, Algorithmica 56, pp. 3-16, 2010（査読有り）

③ Yoshio Okamoto, Ryuhei Uehara and Takeaki Uno, “Counting the Number of Matchings in Chordal and Chordal Bipartite Graph Classes”, Lecture Notes in Computer Science 6911, pp. 296-307, 2010（査読有り）

④Takeaki Uno, “Multi-sorting algorithm for finding pairs of similar short substrings from large-scale string data”, Knowledge and Information Systems 25, pp. 229-251, 2010.（査読有り）

⑤Shuji Kijima, Masashi Kiyomi, Yoshio Okamoto, and Takeaki Uno, “On listing, sampling, and counting the chordal graphs with edge constraints, Theoretical Computer Science 411, pp. 2591-2601, 2010.（査読有り）

⑥Yasuko Matsui, Ryuhei Uehara and Takeaki Uno, “Enumeration of the Perfect Sequences of a Chordal Graph”, Theoretical Computer Science, vol. 411, pp. 3635-3641 (2010). (査読有り)

⑦ Shin-ichi Nakano, Ryuhei Uehara and Takeaki Uno, “A New Approach to Graph Recognition and Applications to Distance-Hereditary Graphs”, Journal of Computer Science and Technology, vol. 24, pp. 517-533. 2009 (査読有り)

⑧ Takeaki Uno and Hiroki Arimura, ″Ambiguous Frequent Itemset Mining and Polynomial Delay Enumeration″, Lecture Notes in Artificial Intelligence 5012, pp. 357-368, 2008 (査読有り)

⑨Takeaki Uno, ″An Efficient Algorithm for Finding Similar Short Substrings from Large Scale String Data″, Lecture Notes in Artificial Intelligence 5012, pp. 345-356, 2008 (査読有り)

⑩Hiroki Arimura, and Takeaki Uno, Mining Maximal Flexible Patterns in a Sequence, Lecture Notes in Computer Science, Springer, 4914/2008, pp. 307-317, 2008 (査読有り)

〔その他〕
ホームページ等
http://research.nii.ac.jp/~uno/codes-j.html

６．研究組織
(1)研究代表者：宇野 毅明 (UNO TAKEAKI)
国立情報学研究所・情報学プリンシプル研究系・准教授
研究者番号：00302977

(2)研究分担者
なし

(3)連携研究者
なし