様式 C-19

科学研究費補助金研究成果報告書

平成２１年５月２６日現在

研究種目：基盤研究(C)
研究期間：2006～2008
課題番号：18500012
研究課題名（和文）　巨大情報のアルゴリズム的超圧縮技術の研究

研究課題名（英文）　Research on techniques for algorithmic super-compression of huge data

研究代表者
伊藤　大雄（ITO HIRO）
京都大学・大学院情報学研究科・准教授
研究者番号：50283487

研究成果の概要：指数爆発あるいは入力そのものが巨大である等の理由で従来困難とされてきた問題に対し、本来の目的を失うことなく視点を変えることによって効率的に解くことができる技法について研究した。特にグラフに「孤立」の概念を導入して部分グラフを列挙する問題、グラフの連結度に関する性質検査、家系図列挙問題などについて効率的なアルゴリズムを与えた。また、一部の問題については、そのアルゴリズムの効率が、ある意味で限界値であることも示した。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2006 年度	1,200,000	0	1,200,000
2007 年度	900,000	270,000	1,170,000
2008 年度	1,400,000	420,000	1,820,000
年度			
年度			
総　計	3,500,000	690,000	4,190,000

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：グラフ，アルゴリズム，列挙，孤立，性質検査，定数時間アルゴリズム

１．研究開始当初の背景

計算機能力の向上に伴い、世界的に大規模な計算機ネットワークができ、我々は瞬時に世界中の情報を得ることができる時代になった。インターネットとそれに関連するデータベースが我々に与えてくれる情報量は前世紀のそれとは、もはや次元が違うというべき詳細かつ膨大なものである。しかし、いくら情報量が膨大でもそれを適切に扱う方法を知らなければ使い物にならない。例えば、インターネットの初期においては、インターネットは一部の専門家が電子メールやファイル転送に利用していたぐらいであったが、Google などの検索エンジンの出現により格段に便利さが増し、それまで一部の専門家だけのものであったインターネットが一気に一般家庭に浸透し、現在の隆盛をみている。

巨大情報を適切に扱う技術が必要なのはインターネット検索に留まらない。一例をあげれば、バイオの領域においては約３０億個の塩基からなる DNA から有意義な情報を得る研究が行われており、特定の病気の発病あるいはその抑制に関わる遺伝子の特定が試みられている。数学の分野でも、膨大な場合分

けを計算機の助力によって解決する試みが行われるようになってきた。ある性質を満たす全てのものを列挙するという問題は、数学や工学の立場では頻繁におきるが、それらも計算機の助力によって大きな進歩を得ている。さらに、理学・工学に留まらず、人文科学の分野においても、そういった試みは始められている。例えば、最も計算機から遠い分野というイメージのある文学研究の分野においてさえ、和歌データベースの構築、それを使用した計算機による分析などが行われるようになってきている。巨大情報とその扱いの重要性に関しては、平成１７年度に特定領域研究「情報爆発」が発足したことからも重要視されていることが分かる。

２．研究の目的

上であげた各問題に共通して存在するのは、巨大情報からその本質をいかに見抜くか、その見抜いた情報をいかにコンパクトに表現するか、という技術である。例えばインターネット検索は、ウェブページの重要度を、簡便なアルゴリズムから得られるページランクという数値で表現したことの勝利である。近年、ハブとオーソリティの概念に基づく２部クリーク抽出の技法の研究が盛んに行われているのも、同じく重要情報を共有していることの本質がそこにあると見抜いたことによる。計算機による総列挙もやみくもに列挙したのではいかに高速な計算機といえども指数爆発の壁を乗り越えることはできないので、本質を見抜いて探索空間を絞り込む技術が重要となる。

我々はこれらの技術を、巨大情報のアルゴリズム的超圧縮と位置づけた。すなわち、そのままでは扱いかねる巨大な情報から、必要な情報を抜き出し、コンパクトに、場合によっては対数的に縮小して表記し、必要に応じて、そこからアルゴリズムによって情報を取り出すという技法である。本研究の目的はこの技術の開発と高度化である。

３．研究の方法

いくつかアプローチ方がある。代表的な物は､(a)定義を見直す方法と(b)指数爆発を逆用して対数的に縮小する方法である。以下でこれらを具体的に説明する。

(a) 定義を見直す方法：ウェブグラフとはインターネットの html ファイルやサーバなどを節点に、リンクを枝に置き換えて作ったグラフのことである。インターネットの検索は、キーワード検索の他にウェブグラフの構造から検索する方法があり、最近はそちらの研究が盛んである。その代表は、クリーク検索である。クリークとはグラフ理論の用語で、全ての節点間に枝が存在する部分グラフのことであるが、ウェブグラフにおいて、似た情報を持ったページはお互いにリンクを張っている可能性が高いという予測から、クリークを検出することによって、関連性の高いサイト群を検出しようという試みである。しかし一般にクリークに関する問題は困難である。最大クリーク問題は代表的な NP 完全問題であり、さらに W[1]-hard 性や、近似比が $\Omega(n^{1-\varepsilon})$であることなどが分かっており､効率的解法に関しては絶望的な状況である。従って研究方向としては、極大クリークの列挙の方向に向かっており、解一つあたり多項式時間で列挙するアルゴリズムがいくつか提案されている。しかし極大クリークの数は指数個存在しうるという事実が分かっており、極大クリークの総列挙は、問題の定義そのものに指数爆発の危険性を内包している。

我々はここで何故クリークが必要なのかを問い直し、その結果、クリークの重要な性質として「内部のリンクが密である」ということの他に「外部とのリンクが疎である」という性質があることに目をつけ、「孤立クリーク」という概念を導入した。k 個のノードからなるクリーク C が孤立しているとは C とそれ以外との間のリンク数が ck 本未満であることを言う。ここに c は孤立度を表すパラメータであり､「c-孤立」などという使い方をする。その結果、任意の定数 c に対して、すべての c-孤立クリークを線形時間で列挙するアルゴリズムを提案することができた。この様に、必要なものの本質を失わずに定義をわずかに変更するだけで、困難な問題を高速に解けるようにすることがこのアプローチである。

(b) 指数爆発を逆用して対数的に縮小する方法：多くの組合せ問題を困難にしている主要因は何といっても「指数爆発」にある。例えば前述の様に、極大クリークはグラフのサイズの指数個あるので、それを一つ一つチェックしてくことは実際問題として不可能であることなどである。しかしこれは逆に見れば小さい記述で膨大な情報を記述できる可能性を示唆している。うまくいけば対数的に情報を圧縮できる可能性が有る。列挙アルゴリズムの研究動向はすべて解を一つ一つ書き出していくのを原則としているが、この対数圧縮を用いた解を出力しておけば、後は非常に簡便なアルゴリズムで解を必要に応じて取り出すことが可能となる。

４．研究成果

主な成果は以下の３点である。

(I) 孤立部分グラフの列挙

孤立の概念を擬クリークおよび二部クリークに拡張した。まず前者ではグラフ G=(V,E)の誘導部分グラフ C で、C に含

まれる節点の（C における）平均次数がα以上で最小次数がβ以上であるようなものを擬クリーク PC(α,β)と定義し、孤立擬クリーク列挙アルゴリズムについて考察した。擬クリークはクリークの一般化になっているので、クリークの列挙以上の難しさが有るのは自明である。本研究によって得られた代表的な結果は以下の通りである。

(1) 任意の定数 $\varepsilon>0$ に対し、1-孤立擬クリーク $PC(k-(\log k)^{1-\varepsilon}, k/(\log k)^{1+\varepsilon})$ を超多項式個含むようなグラフが存在する(従って多項式時間全列挙は原理的に不可能である)。

(2) 任意の定数 c>0 に対して、c-孤立擬クリーク PC(k-log k, k/log k)を全列挙する多項式時間アルゴリズムを与えた。

上記の結果は、多項式時間列挙のある種の限界値を得たことになり、非常に重要な成果であり、アルゴリズムの分野の最高の論文誌である ACM Transactions on Algorithms に採録が決定した[1]。

次に二部クリークについては、以下の結果が得られた。

(1) 1-孤立な２部クリークを指数個含むようなグラフが存在する。

(2) 任意の定数 $\gamma>0$ に対し、二つの部の大きさの比が定数であるような 1-孤立二部クリークを全列挙する多項式時間アルゴリズムを与えた。

この結果はある種の二部クリークの列挙も効率的にできることを示している。

(II) 性質検査

ウェブグラフ等、入力のサイズそのものが膨大なものの場合、その全てでは無く、一部のデータを見るだけで、その性質を判定できれば、非常に便利である。性質Ｐ（Ｐはなんでも良い）と正定数 $\varepsilon>0$ に対し、与えられたグラフＧがＰを満たすならば確率 2/3 以上で受理し、Ｐより遠い（ε 遠隔; 定義は[4]を参照）ならば確率 2/3 以上で拒否するアルゴリズムを性質Ｐの検査アルゴリズムと言い、多くの重要な性質に対し、グラフのサイズに無関係な定数時間で動作する検査アルゴリズムが知られ始めている。グラフの連結性判定に関しては、これまで無向グラフの k-枝連結性の定数時間検査アルゴリズムが知られているのみであった。

これに対し、我々は、まず「有向」グラフの k-枝連結性の判定問題に対して定数時間検査アルゴリズムを構築した[3]。無向グラフの k-枝連結性の検査の場合には、「与えられたグラフが k-枝連結から十分遠ければ連結成分が多数（線形個）存在する」ということは比較的容易に導け、これを元に検査アルゴリズムを構築できる。しかし有向グラフの場合は、この部分は非自明である。我々はさらに、無向グラフの k-「点」連結性の判定問題についても定数時間検査アルゴリズムを得た。点連結を扱うことは枝連結よりも一般に困難であり、本問題についても事情は同じであった。本アルゴリズムの構築のためには、点連結度を下げずに次数を減少させる技法が必要であり、そのために連結度増大問題の基本技法である枝分割(edge splitting)の技法を用いることで、これを解決した。後者のアルゴリズムは計算機科学分野における欧州の最高の会議である ICALP2008 に受理された[4]。

(III) 家系図列挙

個体間の遺伝的距離を入力として、可能な家系図を総列挙する問題を考えた。グラフの距離をそのまま個体間の距離と考えた距離行列の定義の下で行い、この問題の解法を研究した。本問題は解の個数が指数個存在しうるが、対数的圧縮の技法を用いることによって、線形サイズの解の出力で、それら全てを表現する方法を考え、それを求める多項式時間アルゴリズムを与えた[8]。

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１１件）全て査読有り

(1) Hiro ITO and Kazuo IWAMA, Enumeration of isolated cliques and pseudo-cliques, ACM Transactions on Algorithms. (to appear)

(2) Hiro ITO, Mike PATERSON, and Kenya SUGIHARA, Multi-Commodity Source Location Problems and the Price of Greed, Journal of Graph Algorithms and Applications, Vol. 13, No. 1, 2009, pp. 55--73. (to appear)

(3) Yuichi YOSHIDA and Hiro ITO, Testing k-edge-connectivity of digraphs, Journal of System Science and Complexity. (to appear)

(4) Yuichi YOSHIDA and Hiro ITO, Property testing on k-vertex connectivity of graphs, Proceedings of the 31th International Colloquium on Automata, Language and Programming (ICALP 2008), LNCS, #5125, Springer, 2008, pp. 539--550.

(5) Sergey BEREG and Hiro ITO,

Transforming graphs with the same degree sequences, Proceedings of Kyoto International Conference on Computational Geometry and Graph Theory --- in honor of Jin Akiyama and Vasek Chvatal on their 60th birthdays, LNCS #4535, pp. 25--32, 2008.

(6) Kenya SUGIHARA and Hiro ITO, Maximum-cover source location problems with objective edge-connectivity three, Mathematical Methods of Operations Research, Sept., 2008, DOI: 10.1007/s00186-008-0266-1.

(7) Hiro ITO, Mike PATERSON, and Kenya SUGIHARA, Multi-commodity source location problems and price of greed, Proceedings of the Second Workshop on Algorithms and Computation (WALCOM 2008), LNCS, # 4921, Springer, 2008, pp. 169--179.

(8) Hiro ITO, Kazuo IWAMA, and Takeyuki TAMURA, Inferring pedigrees from genetic distances, IEICE Transactions, Vol. E91-D, No. 2, 2008, pp. 162--169.

(9) Sergey BEREG and Hiro ITO, Transforming graphs with the same graphic sequences, Kyoto International Conference on Computational Geometry and Graph Theory --- in honor of Jin Akiyama and Vasek Chvatal on their 60th birthdays, June 11--15, Kyoto, Japan, 2007.

(10) Hiro ITO, Gisaku NAKAMURA, and Satoshi TAKATA, Winning ways of weighted poset games, Akiyama-Chvatal Festschrift, Supplement of "Graphs and Combinatorics," 2007, pp. 291--306.

(11) Hiro ITO, Midori KOBAYASHI, and Gisaku NAKAMURA, Semi-distance codes and Steiner systems, Akiyama-Chvatal Festschrift, Supplement of "Graphs and Combinatorics," 2007, pp. 283--290.

〔学会発表〕(計１１件)

(1) Hiro ITO and Masakazu KADOSHITA, FPT algorithm for the Hamiltonian circuit problem on unit disk graphs parametrized by the area, OR50 Annual Conference (OR50), Sept 9--11, Univ. of York, York, UK, 2008.

(2) 伊藤大雄, 無秩序の代償 (price of anarchy) の理論入門, 第5回日本OR学会中部支部シンポジウム「インターネット時代のゲーム理論」, 2008/09/05, 第二豊田ビル（西館）8階第1会議室（名古屋）.

(3) 伊藤大雄, ２１世紀のネットワーク最適化, ORセミナー「最適化とその実用」, 2008/08/28, 構造計画研究所 本所新館（東京）.

(4) 吉田悠一, 伊藤大雄, 無向グラフのk枝連結性の検査, 信学技報, COM2008-22 (2008-06), pp. 49--55（電子情報通信学会コンピューテーション研究会(COMP), 2008/06/16, 北陸先端科学技術大学院大学）.

(5) Yuichi YOSHIDA and Hiro ITO, On k-connectivity testing in degree-bounded graphs, The First Annual Meeting of Asian Association for Algorithms and Computation (AAAC08), April 26--27, Hong Kong, 2008.

(6) Yuichi YOSHIDA and Hiro ITO, Constant time k-connectivity testing, Chinese Academy of Sciences - Kyoto University Joint Workshop on Mathematical Methods for Informatics, Engineering and Management, Beijing, China, March 17--18, 2008.

(7) Yuichi YOSHIDA and Hiro ITO, Testing k-connectivity on Degree Bounded Graphs, Research Seminar on Connectivity of Graphs and Its Applications, March 10--12, SATAKE Memorial Hall, Hiroshima University, 2008.

(8) Hiro ITO, Mike PATERSON, and Kenya SUGIHARA, Multicommodity source location problems and price of greed, FIT2007, 2007/09/05--07, 中京大豊田キャンパス.

(9) 吉田悠一, 伊藤大雄, 有向グラフにおける k 枝連結性の検査 信学技報, COM2007-18 (2007-06), pp. 17--23（電子情報通信学会コンピューテーション研究会(COMP), 2007/06/29, 北大）.

(10) 宮川博光, 伊藤大雄, 岩間一雄, 部の大きさの比が高々定数倍の孤立２部クリークの列挙, 信学技報, COM2007-18 (2007-06), pp. 9--16（電子情報通信学会コンピューテーション研究会(COMP), 2007/06/29, 北大）.

(11) Hiro ITO, Mike PATERSON, and Kenya SUGIHARA, Multicommodity source location problems and price of greed, Research Seminar on Connectivity of Graphs and Its Applications, Jan 31--Feb 2, SATAKE Memorial Holl, Hiroshima University, 2007.

〔図書〕（計０件）

〔産業財産権〕
○出願状況（計０件）
○取得状況（計０件）

〔その他〕
特になし

６．研究組織
(1)研究代表者
伊藤　大雄（ITO HIRO）
京都大学・大学院情報学研究科・准教授
研究者番号：50283487

(2)研究分担者
岩間　一雄（IWAMA KAZUO）
京都大学・大学院情報学研究科・教授
研究者番号：50131272

(3)連携研究者
中村　義作（NAKAMURA GISAKU）
静岡県立大学・名誉教授
研究者番号：20109200

福田　宏（FUKUDA HIROSHI）
北里大学・一般教養部・准教授
研究者番号：70238484