

「Science と Engineering をつなぐ『Art』を求めて」

ERATO 湊離散構造処理系プロジェクトシンポジウム

湊 真一

(独) 科学技術推進機構 ERATO 湊離散構造処理系プロジェクト*

概要

近年、論理や集合のような基本データ構造を効率よく処理する「BDD」「ZDD」と呼ばれるデータ構造とアルゴリズムが様々な分野で活用されている。このような技法をベースとして、種々の離散構造を統合的に演算処理する技法を体系化し、分野横断的かつ大規模な実問題を高速に処理する技術基盤を構築することを目標として、「ERATO 湊離散構造処理系プロジェクト」が 2009 年 10 月に JST により採択された。今回のイベント企画では、研究総括らによるプロジェクト概要説明の後、ERATO ポスドク研究員、共同研究者、連携研究者による最新の研究成果の紹介をポスターセッション形式で開催し、参加者による自由な技術討論を行う。本シンポジウムにより、情報科学技術分野の多数の研究者に関心を持っていただき、離散構造処理系をベースとした分野横断的な研究者コミュニティの交流を深めたい。以下では、本プロジェクトの概要について述べる。

ERATO プロジェクトとは

JST の戦略的創造研究推進事業には 3 種類あり、牧場型の若手育成を目指す「さきがけ」、八ヶ岳のような山脈型の分野形成を目指す「CREST」、そして富士山型の突出した研究を強力に推進するのが「ERATO」である。この中では ERATO が最も伝統があり（昭和 56 年発足）、科学技術の新しい潮流を形成し、新技術の創出に資するシーズを生み出すことを狙いとしている。採択件数は年間 4～5 件で、これまでに約 100 人の研究者が選ばれている。制度発足当初は、物理・化学・生命科学の分野から選ばれていたが、10 年程前から、情報分野も 1～2 年に 1 件程度採択されるようになった。プロジェクトの期間は 5 年半、人件費や設備費を含む研究費総額は多くの場合 10～15 億円程度、人員規模は 10～15 人程度となっている。

ERATO は選考方法にも特色がある。まず他薦と JST 独自調査により 1000 人規模の母集団リストを構成し、その後、分野ごとに 10 名程度の一次候補者をノミネートする。この段階で初めて本人に通知し、研究構想の応募を依頼する。候補者は突然の連絡に驚き悩まされる。（例年 2～3 割は応募辞退している。）候補者に選ばれたことは基本的には本人以外には知らされない。研究テーマは、国の重点分野のいずれかに該当しさえすれば、内容は全く自由で、背景、実施体制、期待される成果など、全ての構想を英文で表現することを求められる。その後、さらに絞込みが行われ、外国人審査員を交えた最終面接を経て、各分野で 1 人だけが選出される。

以上のように、ERATO は日本の大型プロジェクトでは最も伝統があり、選考プロセスにも多大な時間と人手がかかっている。さらに ERATO では、分野・組織にとらわれずに機動的なプロジェクトを構成することを狙いとしているため、大学への助成金や委託研究ではなく、JST の直轄事業として、本務の研究室とは独立したオフィスを設置し、研究員やスタッフも原則 ERATO 予算で直接雇用する制度となっている。

*札幌市北区北 14 条西 9 丁目 北海道大学 大学院 情報科学研究科 工学系 C306 号室 phone: 011-728-8270
email: minato@ist.hokudai.ac.jp, www: <http://www-erato.ist.hokudai.ac.jp/>

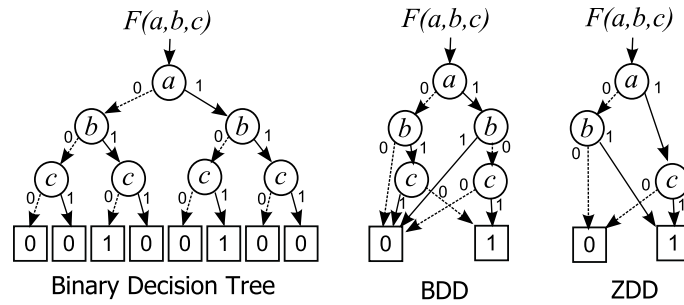


図 1: 場合分け二分木, BDD, ZDD.

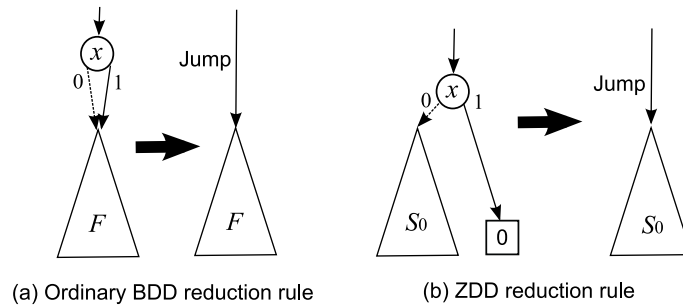


図 2: ZDD の簡約化規則.

本プロジェクトの技術的背景

離散構造は、離散数学および計算機科学の基礎をなすものであり、集合論、記号論理、帰納的証明、グラフ理論、組合せ論、確率論などを含む数学的な構造の体系である。およそ計算機が扱うあらゆる問題は、単純な基本演算機能を要素とする離散構造として表現することができ、そのような離散構造の処理に帰着される。離散構造の処理は、最終的には膨大な個数の場合分け処理を必要とすることが多い。種々の離散構造データを計算機上にコンパクトに表現し、等価性・正当性の検証、モデルの解析、最適化などの処理を効率よく行う技法は、計算機科学の様々な応用分野に共通する基盤技術として非常に重要である。典型的な応用分野としては、ハードウェア/ソフトウェアの設計、大規模システムの故障解析、制約充足問題、データマイニングと知識発見、機械学習と自動分類、バイオインフォマティクス、web 情報解析、等、多種多様であり、現代社会に対する大きな波及効果を持つ。

離散構造に関する最も基本的なモデルの 1 つとして論理関数がある。論理関数を計算機上で処理する工学的技法に関しては、古くは 1938 年に Shannon がブール代数による論理回路の設計法を示したことに始まり、それ以降現在に至るまで長い歴史がある。論理関数のデータ構造としては、当初は真理値表表現や積和/和積標準形 (DNF/CNF) が用いられてきたが、1986 年に R. E. Bryant が画期的な論文 [1] を発表して以降、BDD を用いた論理関数の処理技術が急速に発展した。

BDD (Binary Decision Diagram; 二分決定グラフ) は、VLSI 設計の分野で考案された論理関数データのグラフによる表現である。これは、論理関数の値を全ての変数について場合分けした結果を図 1 のように二分木グラフで表し、これを簡約化することにより得られる。このとき、場合分けする変数の順序を固定し、(1) 冗長な節点を削除する、(2) 等価な節点を共有する、という処理を可能な限り行うことにより「既約」な形が得られ、論理関数をコンパクトかつ一意に表せることが知られている。BDD によるデータ圧縮率は、論理関数の性質にも依存するが、例題によっては数十倍～数百倍以上もの圧縮率が得られる場合がある。また、2つの BDD を入力とし、それらの二項論理演算 (AND, OR 等) の結果を表す BDD を直接生成するアルゴリズムが知られており、ハッシュテーブルを巧みに用いることで、データが計算機の主記憶に収まる限りは、その記憶量にほぼ比例する時間内で論理演算を効率よく実行できるという特長を持っている。

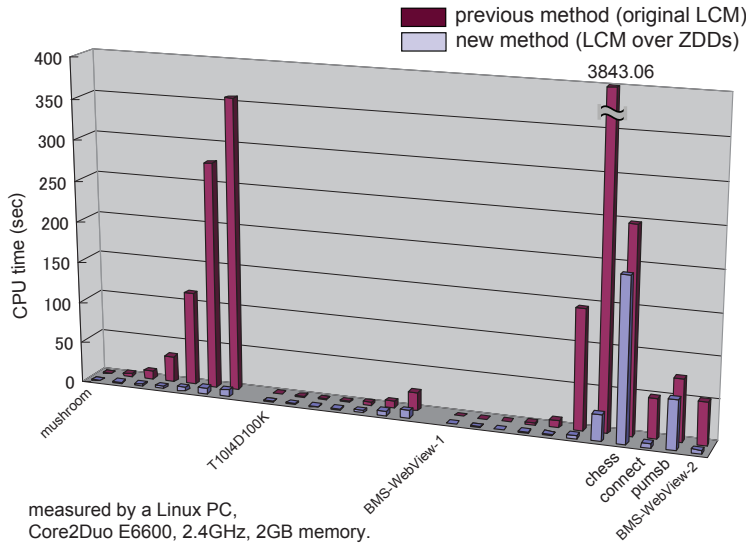


図 3: LCM over ZDDs の実験結果.

BDD はオンメモリの大規模論理データ処理系であり、メモリにランダムアクセスできる計算機モデルの能力を最大限に引き出す技法とも言える。近年の PC の主記憶の大規模化により、従来は記憶容量不足で扱えなかった規模の問題が、現実的に扱えるようになったため、特に 2000 年以降、BDD の応用範囲が拡大している。

BDD は、元々は論理関数の表現法であるが、 n 個の要素からなる集合の各要素を、 n 変数の論理関数の入力変数に対応付けることにより、集合族データを表現することもできる。BDD 表現は、類似する部分集合が多く含まれるような集合族データを表現する場合に顕著なデータ圧縮効果が得られ、高速な演算処理が可能となる。このような性質から、BDD は、VLSI 論理設計の分野だけではなく、様々な制約充足問題の解法や、大規模システムの故障解析など、0/1 の論理の組合せや集合を扱う多種多様な問題に応用されるようになった。

このような集合族データの処理に特化した BDD として、湊が 1993 年に考案・命名したものが ZDD (Zero-suppressed BDD; ゼロサプレス型 BDD) である [3]。ZDD は、通常の BDD と異なる簡約化規則を持つ。すなわち、図 2 に示すように、1-枝が 0-終端節点を直接指している場合に、この節点を取り除く。その代わり、通常の BDD で削除されるような節点は削除しない。この簡約化規則は、特に疎な集合の族を表現する際に顕著な効果がある。例えば、集合要素の平均出現頻度が 1% であれば、ZDD は BDD より 100 倍コンパクトになる可能性がある。世の中で扱われる実問題では、集合要素の平均出現頻度が非常に小さい場合が多い（例えば、店舗での陳列商品の総数に比べて、顧客が 1 度に購入する商品数は極めて少ない）ため、ZDD による圧縮効果は極めて大きい。

ZDD は、BDD の改良技術として世界的にも広く定着しており、BDD と並んで標準的に使われている。特に、最近では、データマイニング・知識発見への応用分野で ZDD の有効性が示されている。その一例として、頻出パターン抽出問題への応用が挙げられる。筆者らは、当時の世界最高速のアルゴリズムであった「LCM」に、さらに ZDD 技術を導入した「LCM over ZBDDs」を開発した [4]。図 3 に示すように、例題によっては数十倍～数百倍にも及ぶ圧倒的な性能改善が見られる。特に、解として抽出されたパターンの個数が膨大になる場合に顕著な効果がある。

一般に、頻出パターンマイニングでは、頻度の閾値を高く設定し過ぎると自明なパターンしか見つからず、逆に閾値が低すぎると抽出パターン数が膨大になり過ぎて、興味あるパターンの発見が困難になる。したがって、ちょうど良い適切な頻度を設定する必要があるが、例題に応じて実験的に決めるしかなく、それが本当に適切な値だったかどうか確かめようがないという問題があった。これに対して ZDD を使えば、互いに類似する膨大な個数のパターンを

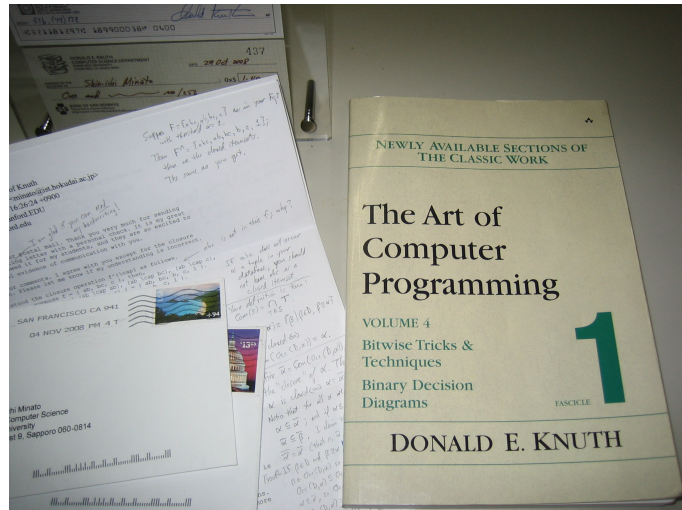


図 4: Knuth 先生からの手紙と掲載された本

ひとまとめにしてコンパクトに圧縮して、計算機の主記憶上に表現することができる。さらに、抽出したパターン全体に対して ZDD の集合演算を使って一定の制約条件を加えて、効率よく絞込みを行うことができる。これは従来の明示的な表現方法では、結果を出力するのにも膨大な時間がかかり、たとえ出力できたとしても、その後に意味のある解析処理を現実的な時間で行うことは不可能であった。つまり、ZDD を使うことにより、これまで扱えなかった超大規模な頻出パターンを、現実解析できるようになった訳である。BDD は当初は LSI 設計自動化の分野で考案・発展してきたものであるが、ZDD を始めとする BDD の進化形がいくつか考案され、広く計算機科学全般に応用範囲が広がっている。

本プロジェクト提案に至った契機

2008 年 9 月に、D.E.Knuth から湊に突然のメールが届いた。彼が執筆中の名著「The Art of Computer Programming」[2] の最新の章は専ら BDD について 140 ページに渡って書かれており、その中で ZDD についても詳しく執筆したいので、ZDD の考案者として、校正作業に協力して欲しいとの依頼であった。世界的に有名なこの教科書に、日本人が考案したデータ構造が項目として詳細に掲載されるのは、今回が初めてのことである。

BDD の改良技法はこれまでに多数提案されているが、Knuth はその中で特に ZDD を重要視し、章の後半部分 30 ページ以上を割いて、ZDD の解説や演習問題を載せている。(それまで私は ZBDD と呼称していたが、BDD と同格であるべきとの Knuth の指摘により ZDD に短縮された。) 本書で Knuth は、ZDD の基本演算を整理し直し、集合族を扱う代数系を定義し「Family algebra」と命名された。これにより BDD/ZDD が離散構造の基本処理系として汎用的に活用できることが、より明確になった。(図 4)

本書には BDD に関する 264 問もの演習問題と解答が含まれている。しかし、これで全てをカバーし尽くしているわけではない。本書ではコンピュータプログラミングの観点で汎用的に重要な技術を徹底的・網羅的に扱っているが、特定の応用向けの技術は範囲外として記載していない。さらに、当然ながら、本に記載されるのは執筆時点までのことであり、未来に考案される事項は記載されない。本章の校正作業がほぼ終わった頃に、Knuth 氏は ZDD 演算に関するちょっとしたアイデアを思いつかれて、それを湊あてに送って来られた。しかしながら、Knuth 氏は、自分はすでに高齢であり、健康でいられる間に次の巻の執筆を急がないといけないので、これ以上立ち止まってられない、あとは自由に考えてくれ、と述べて、次の章に進んで行かれた。

以上のような状況から、離散構造を統合的に扱う基本処理系として BDD/ZDD を位置づけ、分野横断的な応用を持つ技術体系として研究開発を進める好機にあると考えた。本分野に人的資源を投入し、国際的競争力を持つ研究者集団を形成することは、我が国の研究戦略上も

交流も容易であり、非常に優れた研究環境にあると言える。さらに東京地区・関西地区にサテライト拠点を設置し、TV 会議システムによる密な意思疎通を行いながらプロジェクトを推進している。東京地区は東工大の GCOE プログラム「計算世界観」との共同研究を実施し、大岡山キャンパス内にサテライトラボを置かせて頂いている。関西地区は京阪神の中心である大阪・梅田駅付近にオフィスを借りて設置しており、梅田地区に集積する企業との連携および関西に分散する大学の研究者との連携を意図している。なお、東京、関西の直轄研究拠点以外にも、例えば九州地区にも TV 会議拠点を置かせてもらう方向で調整中である。

各研究拠点に対応して、以下のようなサブグループ組織を構成しており、それぞれについて、極めて有力な若手中堅の研究者にグループリーダーを委嘱している。

- (1) 離散構造基盤グループ (GL: 研究総括兼任)
札幌メインオフィス (研究員・技術員 10 名程度)
- (2) 機械学習・制約充足応用グループ (GL: 産総研・津田宏治主任研究員)
東京地区 (研究員・技術員 若干名)
- (3) 統計・マイニング応用グループ (GL: 阪大・鷲尾隆教授)
関西サテライトオフィス (研究員・技術員 若干名)

研究拠点は上記の通り 3 箇所設置しているが、フルタイム雇用の若手ポスドク研究員は、なるべくメインの札幌に集めて、活性化・相乗効果を図ることとしている。一方、各地区のサテライト拠点では、各分野の第一線で活躍中の大学教員や企業研究者との間で、共同研究等の連携を積極的に行い、メインの札幌地区との間で有機的なネットワークを形成することを狙っている。

おわりに

以上、ERATO プロジェクトの概要について紹介した。本プロジェクトは 2009 年 10 月に発足し、半年間の準備期間を経て、2010 年 4 月より本格的な研究活動を開始した。プロジェクトはこれまでのところ比較的順調に立ち上がって来ており、今後は具体的な研究を進め、地道に成果を積み上げていくフェーズに入っていく。

ソフトウェア系の ERATO は、特殊で高価な実験装置はあまり必要ではない。優れた研究者集団を形成するための環境作りが最も重要である。本プロジェクトのパネルオフィサー（審査責任者）の西尾章治郎先生からは「札幌を、優秀な人材が集まる梁山泊のようにして欲しい」との激励を頂いている。東京・関西の拠点と活発な研究交流を行いながら、優秀な若手人材を見出し、高いレベルの「梁山泊」を形成し、一定期間じっくりと研究を行い、コアとなる「Art」を形成し、学界・産業界に還流させていきたいと考えている。関係各位のご理解・ご協力をお願いする次第である。

参考文献

- [1] R. E. Bryant. Graph-based algorithms for Boolean function manipulation. *IEEE Transactions on Computers*, C-35(8):677–691, 1986.
- [2] D. E. Knuth. *The Art of Computer Programming: Bitwise Tricks & Techniques; Binary Decision Diagrams*, volume 4, fascicle 1. Addison-Wesley, 2009.
- [3] S. Minato. Zero-suppressed BDDs for set manipulation in combinatorial problems. In *Proc. of 30th ACM/IEEE Design Automation Conference*, pages 272–277, 1993.
- [4] S. Minato, T. Uno, and H. Arimura. LCM over ZBDDs: Fast generation of very large-scale frequent itemsets using a compact graph-based representation. In *Proc. of 12th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2008)*, (LNAI 5012, Springer), pages 234–246, 5 2008.