# 大規模データ処理のための離散構造処理系

奈良先端科学技術大学院大学・助教　川原　純

様々な情報源から得られたビッグデータを活用するためには、それらをストレージに格納しなければならないが、ストレージの容量には限りがあるため、なるべく圧縮された形でデータを保持したい。しかしながら、例えば gzip のような圧縮形式を用いると、データの検索、サンプリング、フィルタリング等のデータの活用時に、圧縮されたデータを逐一展開しなければならず、時間がかかり過ぎる。近年、圧縮された状態のデータを展開せずに活用できるデータ構造が提案されており、特に、簡潔データ構造や、二分決定図等のアルゴリズム的手法が盛んに研究されている。本講演では、後者の二分決定図を用いたデータの処理系について解説を行う。

ゼロサプレス型二分決定図（Zero-suppressed Decision Diagram, ZDD）は、集合族を表現するためのデータ構造である。例えば、スーパーマーケットの客の購買データを記憶するのに ZDD を用いることができる。アイテムを $x_1, x_2, \ldots$ と表すとすると、客の購買データは $\{x_1, x_2, x_5\}$, $\{x_1, x_3, x_4, x_7\}$ のように集合で表され、データベース全体は集合族となる。ZDD によって、集合族を効率的に保持できるだけでなく、様々な集合演算を ZDD の形のまま高速に行うことが可能である。具体的には、2 つの ZDD が与えられたとき、それらの ZDD が表す集合族の和集合や共通部分等、多くの集合を求める演算ができる（図 1）。また、条件を指定してのフィルタリングや一様サンプリング等が可能である。これらの演算は、ZDD のライブラリを用いることで簡単に記述できる。

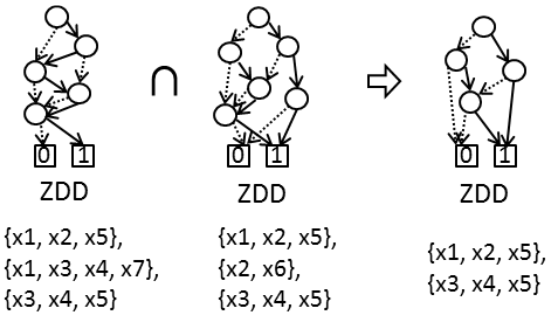


図 1　ZDD 同士の演算

ZDD を用いて、グラフ上の 2 点間のすべてのパスを列挙するアルゴリズムが、Knuth の著書 The Art of Computer Programming 4A で紹介されている。Knuth のアルゴリズムを用いると、例えば、15 × 15 グリッドグラフ上の左上隅から右下隅に至るパス 227449714676812739631826459327989863387613323440（$= 2.27 \times 10^{47}$）本を、ZDD の形で圧縮された状態で出力できる。計算時間はわずか数分である。出力は ZDD の形であるため、条件を満たすパスの抽出やサンプリング等が容易である。本研究では、Knuth のアルゴリズムを基に、パスだけでなく、全域木や頂点被覆等、様々な部分グラフを表す ZDD を構築できるアルゴリズムを開発し、それをフロンティア法と名付けた。本講演では、フロンティア法の応用事例として、通信ネットワークの信頼性評価や、配電網の構成法についても紹介する。また、大規模なグラフの集合を扱うソフトウェア、graphillion（http://graphillon.org、井上 武氏制作）を紹介する。