



HOKKAIDO UNIVERSITY

Title	Studies on Decision Diagrams for Efficient Manipulation of Sets and Strings [an abstract of dissertation and a summary of dissertation review]
Author(s)	伝住, 周平
Degree Grantor	北海道大学
Degree Name	博士(情報科学)
Dissertation Number	甲第11749号
Issue Date	2015-03-25
Doc URL	https://hdl.handle.net/2115/58737
Rights(URL)	https://creativecommons.org/licenses/by-nc-sa/2.1/jp/
Type	doctoral thesis
File Information	Shuhei_Denzumi_review.pdf, 審査の要旨



学位論文審査の要旨

博士の専攻分野の名称 博士 (情報科学)

氏名 伝住 周平

審査担当者 主 査 教 授 湊 真一

副 査 教 授 Zeugmann Thomas

副 査 教 授 有村 博紀

学位論文題名

Studies on Decision Diagrams for Efficient Manipulation of Sets and Strings

(集合および文字列を効率よく操作するための決定グラフに関する研究)

近年の高速ネットワークと大容量記憶装置の急速な発展によって、データベースに蓄積された大量のデータを解析処理し、自明でない有用な規則性やパターンを抽出するデータマイニングの技法が進展し、社会の様々な場面で活用される重要な基盤技術となっている。このような大規模データベースの解析処理を行うためには、論理、集合、文字列などの基本的な離散構造をできるだけ少ないメモリ量で表現し、効率よく処理する技術が必要となる。特に、与えられた文字列データを整理・圧縮して格納する索引化の技法は、これまでもトライ、接尾辞木、接尾辞配列など、多くの技法が考案されている。索引の応用においては、メモリ効率や検索速度だけではなく、データの挿入・削除・統合などの高度な操作を行うための処理性能も重視される場合があり、現在も、理論と実用の両面から活発な研究開発が行われている研究分野である。

著者は、組合せ集合データを効率よく圧縮索引化する「ゼロサプレス型二分決定グラフ」(ZDD: Zero-suppressed Binary Decision Diagram), およびそれを文字列処理に拡張した「系列二分決定グラフ」(SeqBDD: Sequence Binary Decision Diagram) に関する研究を行ってきた。SeqBDD は 2009 年に Loekit らにより考案された比較的新しいデータ構造で、任意の有限長文字列の集合を表現でき、文字列の検索だけでなく、挿入・削除・統合などの動的な操作に優れた技法であるが、理論的性質や実用性能など未解明の部分が多く残されていた。本論文の前半 (3 章, 4 章) は、SeqBDD に関する著者の理論的および技術的な貢献をまとめたものである。3 章では、SeqBDD の理論的性質について述べており、具体的には、最小な SeqBDD と最小な非巡回決定性有限オートマトン (ADFA) のサイズの理論的関係性、および SeqBDD の構築にかかる時間・空間計算量を明らかにしている。続く 4 章では、SeqBDD を現実の情報検索処理に応用するための実用的手法を提案している。与えられた文字列が含む部分文字列を全列挙した集合を SeqBDD で効率よく表現できることは以前より知られていたが、実応用においては、それぞれの部分文字列の出現位置や頻度を保持することも必要となる。著者は、SeqBDD に出現位置や頻度を保持する機能を追加し、さらに単一の文字列だけではなく、膨大な個数の文字列の集合を SeqBDD によってまとめて与えたときに、それらの部分文字列を全列挙し、出現位置や頻度を保持する技法を提案し、実験により有効性を確かめている。

一方、現実のデータベースを組合せ集合として表現した場合、非常に多数のアイテムの中からごく少数を選び出すような、疎 (sparse) な組合せが大量に発生することが多く見られる。このようなデータベースを ZDD で索引化した場合、ごく少数のアイテムからなる組合せを 1 つ検索するために、ZDD を上から下まで順にたどるため、データベースに出現する全アイテム種類数に比例する計

算時間がかかり効率的でない。さらに ZDD の各節点は数十文字分のメモリを消費してわずか 1 文字しか表さないため、メモリ効率の改善余地が大きい。5 章では、このような問題を解決するために著者が考案した DenseZDD というデータ構造について述べている。DenseZDD は、与えられた ZDD を理論的限界まで圧縮し、さらに任意の組合せの所属の有無の判定を、組合せのサイズに比例する時間で実行できるという特長を持つ。ただし DenseZDD は、挿入・削除・統合などの動的な操作は提供しないので、従来の ZDD と DenseZDD を組合せたハイブリッドな処理系を構成する技法を提案している。なお、DenseZDD の圧縮技法は、文字列を扱う SeqBDD にも拡張が可能である。

本論文の成果は、次のようにまとめられる。

1. SeqBDD に関するいくつかの重要な理論的性質を明らかにした。
2. 大量の文字列集合データの部分文字列を SeqBDD で全列挙し、各部分文字列の出現位置や頻度を保持する手法を提案した。
3. ZDD を理論的限界まで圧縮し、疎な組合せ集合に対する検索時間を高速化した DenseZDD と呼ぶデータ構造を考案した。

これを要するに、著者は、SeqBDD の理論的性質の解明と文字列索引のため実用的な拡張、および ZDD の空間・時間計算量を改善する圧縮技法を提案し、組合せや文字列データの索引の処理性能向上が可能であるという新知見を得たものであり、離散構造処理アルゴリズムとデータベースの理工学において貢献するところ大なるものがある。よって著者は北海道大学博士 (情報科学) の学位を授与される資格あるものと認める。