



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_abstract.pdf, 論文内容の要旨



学 位 論 文 内 容 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

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, World Wide Web (ウェブ) や大容量記憶装置に代表される高速ネットワークの急速な発展によって, 経済・社会活動・科学技術に関する複雑かつ膨大な量の知識が, インターネット上に流通・蓄積されるようになってきた. このように, 爆発的に増大する一次データをそのままの形で扱うことが困難な状況になりつつあり, 何らかの方法で解析処理を行い, データに含まれる有意義な情報を取り出す効率のよい方法の研究開発が緊急の課題となっている. このような背景からデータマイニング技術に関心が集まっている. データマイニング技術とは, 大規模なデータベースに蓄積された大量のデータを解析処理し, 知識として自明でない規則性やパターンを半自動的に取り出す方法についての科学研究である. データマイニング研究は, 1990 年代初頭から顕在化し, 主に定型的な構造を持つ関係データベースを対象に, 理論と応用の両面で活発な研究が進んできた. データマイニング技術は, 物流効率化やウェブ情報検索, 疫学調査, 遺伝情報解析等, 今や社会の様々な局面で必要とされる重要な基盤技術の 1 つとなっている. 計算機科学の分野において離散構造は重要な概念の一つであり, その中には論理, 集合論, 文字列処理といった重要な研究領域が多数含まれている. 計算機上で離散構造をコンパクトに表現し, 効率よく操作する技術はデータマイニングを始めとする多くの問題に共通して必要とされているため, 離散構造を効率的に操作する技術が近代社会に与える影響は極めて大きい. 文字列や集合といった離散構造をまとめてコンパクトに格納する索引は数十年にわたって幅広く研究されてきた. 文字列索引の例を挙げると, トライや有限オートマトン, トランスデューサや接尾辞木, 接尾辞配列, 部分文字列オートマトンなど枚挙にいとまがない. 日々生産されている膨大な量の離散構造データが急速に増加していることにより, コンパクトな索引に対する要求は日増しに大きくなり続けている. 応用において索引が要求されるのは, コンパクト性だけではなく, データの挿入・削除・統合などの高度な操作を効率よく行う能力も求められている.

本論文は, 3 つの章から構成されている. 膨大な系列データを操作することは文字列処理の分野においてもっとも重要な問題の一つである. 第 3 章では, 近年 E. Loekito らによって提案された, 文字列集合を格納し操作する新しいデータ構造である系列二分決定グラフ (SeqBDD) に関して議論を行った. SeqBDD は非巡回決定性有限オートマトン (ADFA) と二分決定グラフ (BDD) の子孫として捉えることのできるデータ構造であり二分決定木に対して簡約化規則を適用することで得られる. 最小 ADFA と同様に文字列集合をコンパクトに表現すると同時に, BDD から受け継いだ集合演算を効率よく実行することができる. SeqBDD は BDD から継承した多様なアルゴリズムを保持しており, 文字列集合に対して効率良く演算を実行することができる. このデータの格納・検索だけにとどまらない豊富な演算体系は, 従来のデータ構造と比べ大きな特徴となっている. SeqBDD は任意の有限系列集合を表現・操作できる点で他の索引構造よりも幅広い用途が考えられる. 本研究では, SeqBDD の基本的な性質を明らかにした. 具体的には, 簡約化規則を適用することによって得られる SeqBDD が最小な SeqBDD と等しくなるということ, 非自明であった最小な

SeqBDD と最小な ADFA のサイズの関係性, 最小化や二項集合演算および与えられた文字列集合に対する SeqBDD の構築に係る時間・空間計算量, を示した. また, 実データおよび人口データを用いた実験によって上記を確認した.

近年の大規模なテキストやシーケンスデータの増加にともない文字列処理技術の重要性も増している. 特に, 部分文字列索引を用いた効率のよいテキストの抽出や解析は生命情報学や自然言語処理, 文字列マイニングなどの分野で大きな注目を集めてきた. 膨大なデータからデータマイニングを行うには, データを索引付けして効率よく保持する技術の利用が不可欠である. 部分文字列索引は情報検索の分野で広く知られており, 実際の情報検索エンジンにも使われている. また, 多くの実応用においては, 頻度と位置情報を保持した索引が必要とされる. 部分文字列索引の構築は文字列処理における最も重要な問題の一つであり, 単一の文字列パターンに関しては配列や簡潔構造に基づいた効率のよい構築アルゴリズムが存在する. この問題の自然な拡張として, 複数の文字列パターン言語の集合和のなすような言語族を考えることができるが, この集合和族に対して効率よく索引を構築可能かどうかは未解決である. 特に, 有効非巡回グラフ (DAG) のみを対象とする索引の研究はほとんどない. 第 4 章では, SeqBDD のフレームワークの中で索引を構築する問題を考える. まず, SeqBDD 上での単一文字列パターンに対する索引構築手続きを開発し, 次に, その対象を SeqBDD で表現された文字列集合に拡張して, 効率のよい索引構築アルゴリズムを実現する方法について研究を行った. 本研究では, 複数の文字列パターンに対する効率よい構築アルゴリズムを実現することができた. これにより指数的な種類の文字列を圧縮して表現している SeqBDD からそれを展開せずに部分文字列索引を構築できる. また, 自然数を 0 と 1 からなる 2 進数で表現しそれらを文字列として SeqBDD 上で扱う手法を提案した. この手法により文字列の入力グラフ上での出現位置情報を付加することができた. この SeqBDD の形式で与えられた DAG に対する索引を SeqBDD-FP として定義した. そして計算機実験によって SeqBDD-FP の有効性を確認した.

実生活に関わる多くの問題において, 集合の集合, すなわち集合族を扱う必要性に直面することがある. 巨大な集合族を操作することは情報抽出や情報統合, そしてデータマイニングの分野における重要な技術の一つである. この目的のために, 特別なタイプの BDD であるゼロサプレス型二分決定グラフ (ZDD) が用いられることがある. しかしながら, 現在の技術では ZDD を主記憶上に保持するのに大量の領域が必要であり, さらに ZDD を集合族の索引として使用した際に実行される集合の所属性判定に要する時間計算量も大きいという問題がある. 第 5 章では, ZDD をさらに圧縮し静的な索引とする DenseZDD というデータ構造を導入する. 本手法では, 簡潔データ構造と呼ばれる技術を用いることで, 集合族をコンパクトに表現するだけでなく, 所属性判定も高速に実行できるようにした. さらに静的な演算だけではなく動的な演算である集合演算も可能となるように, DenseZDD と従来の ZDD のハイブリッド手法を提案した.

本研究を通して, 集合および文字列を操作する決定グラフを研究した. 文字列を扱う SeqBDD の基本的な性質が明らかになったことで, SeqBDD の研究を発展させていく基盤が整った. さらに, SeqBDD を用いた DAG に対する索引は遺伝子配列データや二次元軌跡データなど幅広い応用が考えられる. 簡潔データ構造によって ZDD を圧縮表現する DenseZDD の技術はそのまま SeqBDD にも適用可能である. 従来のポインタを使用した決定グラフと簡潔データ構造を使用した決定グラフのハイブリッド手法によって, 今まで扱うことが困難だったサイズの問題が解決されることが期待される. 本研究の成果が今後の技術発展に必要な知見の蓄積としても貢献することを願う.