様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　５　月　３１　日現在

機関番号：　１２６０１
研究種目：　研究活動スタート支援
研究期間：　2015～2016
課題番号：　１５Ｈ０６１０１

研究課題名（和文）二分決定グラフに基く非巡回有向グラフ処理アルゴリズムの研究

研究課題名（英文）Research on Algorithms to Process Directed Acyclic Graph Based on Binary Decision Diagrams

研究代表者

伝住　周平（Denzumi, Shuhei）

東京大学・大学院情報理工学系研究科・助教

研究者番号：９０７５５７２９

交付決定額（研究期間全体）：（直接経費）　　2,300,000円

研究成果の概要（和文）：二分決定グラフ(Binary Decision Diagram; BDD)というデータ構造を用い，大規模データベースを巡回の無い有向グラフとして効率良く保持し，圧縮したまま解析処理する手法を開発した．大規模データベースを非巡回有向グラフとして圧縮表現する手法の開発，BDDに基くデータベース解析処理アルゴリズムの効率化の研究，本基盤技術の実データへの応用と性能評価及び課題のフィードバックを実施した．

研究成果の概要（英文）：I have developed methods to process and analyze huge databases that are efficiently represented as directed acyclic graphs, by using Binary Decision Diagrams (BDDs). I have researched data structures to represent huge databases as directed acyclic graphs in compressed form, efficient algorithms to process and analyze databases based on BDDs, and applying these fundamental techniques for real data with performance evaluations and feedbacks.

研究分野：　アルゴリズムとデータ構造

キーワード：　非巡回有向グラフ　二分決定グラフ　アルゴリズム　データ構造　圧縮処理　簡潔データ構造　文字列索引　近似文字列照合

１．研究開始当初の背景

(1)【本研究に関連する国内・国外の研究動向及び位置づけ】

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．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．情報検索に用いられる代表的な索引構造としては部分文字列索引がある．これはテキスト中に現れた単語とその位置を予め記憶しておくことで，その単語の出現場所を簡単に引き出せるようにするデータ構造である．部分文字列索引は広く知られており，実際の情報検索エンジンにも使われている．このほか，Directed Acyclic Word Graph (DAWG)[1]や接尾辞木[2]，接尾辞配列[3]など，現在まで様々な索引構造が研究されている．ただし，従来研究では一系列に対する効率の良いアルゴリズムが重点的に研究されてきたため，複数系列の圧縮同時処理はあまり研究されてこなかった．

(2)【研究代表者の当時までの研究成果を踏まえた着想に至った経緯】

研究代表者はこれまで二分決定グラフの中でも特に文字列処理に秀でた系列二分決定グラフ(Sequence BDD)に関して研究を行い，その基本的な性質の解明や，コンパクト性・高速検索性の向上などを実現してきた．特に，部分文字列索引を構築する際に，従来のほとんどのデータ構造が１本の文字列を入力とするところを，有限長文字列の有限集合を圧縮して表現できる有向非巡回グラフを入力とする SeqBDD に基づく部分文字列索引を提案した．この時に，文字列処理の分野では高速でメモリ効率も良い優れたアルゴリズムが多くあっても，それらを有向非巡回グラフに拡張した研究は非常に少ないことに気がついたことが本研究の着想へと至った経緯である．

(3)【これから発展させる研究内容】

これまでは主に文字列を対象として研究を行ってきたが，集合や順列，論理関数，木といった種々の離散構造を格納する大規模データベースを対象とし，それらを非巡回決定性グラフの形式で表現する手法を開発する．また，SeqBDD を含む二分決定グラフの仲間も有向非巡回グラフの一種と捉えることが可能で，互いの高い親和性を活かした高速なアルゴリズムの開発を行う．二分決定グラフは自身が保持するデータを動的に更新する演算体系を既に保有しているが，それらをさらに発展させて様々な要求に応えることができるようなアルゴリズムを提案する．さらに，複数の研究者と協力し本技術をプライバシー保護等の多様な実問題に適用する方法を考えていく．

[1] A. Blumer, J. Blumer, D. Haussler, A. Ehrenfeucht, M. T. Chen, J. I. Seiferas, The smallest automaton recognizing the subwords of a text, Theor. Comput. Sci., 40, 31-55, 1985.

[2] P. Weiner, Linear pattern matching algorithms, Proc. IEEE 14th Annual Symposium on Switching and Automata Theory, 1-11, 1973.

[3] U. Manber and E. W. Myers, Suffix arrays: a new method for on-line string searches, SIAM J. Comput., 22(5), 935-948, 1993.

２．研究の目的

大規模データベースを巡回の無い有向グラフとして圧縮表現する手法の開発については，既存の二分決定グラフの知識を元にデータに適したデータ構造を開発する．二分決定グラフの仲間には，集合族を表現するゼロサプレス型二分決定グラフ(Zero-suppressed BDD)や順列の集合を表現する順列二分決定グラフ(Permutation BDD)などが提案されている．これらの技術を利用して扱うデータベースを効率良く表現できるようなデータ構造とアルゴリズムの組合せを考案する．

BDD に基くデータベース解析処理アルゴリズムの効率化の研究については，単一の要素を入力としている従来のアルゴリズムの中でも高速なものを非巡回有向グラフ入力に拡張する方法を研究する．中でも部分文字列索引の構築に関しては既存の線形時間構築アルゴリズムの原理を解析することで，多数の文字列を表す非巡回有向グラフ(図．1)に対する索引の高速構築を目指す．

{ε, a, aa, aaa, aaaa, aaaaa, aaaab, aaaaba, aaab, aaaba, aaabaa, aaabab, aaababa, aaabb, aaabba, aab, aaba, aabaa, aabab, aababa, aabb, aabba, ab, aba, abaa, abaaa, abaaaa, abaaab, abaaaba, abaab, abaaba, abaabaa, abaabab, abaababa, abaabb, abaabba, abab, ababa, ababaa, ababab, abababa, ababb, ababba, abb, abba, abbaa, abbab, abbaba, abbb, abbba, b, ba, baa, baaa, baaaa, baaab, baaaba, baab, baaba, baabaa, baabab, baababa, baabb, baabba, bab, baba, babaa, babab, bababa, babb, babba, bb, bba, bbaa, bbab, bbaba, bbb, bbba}

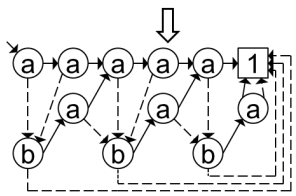

図 1. 非巡回有向グラフによって文字列集合が圧縮して表現される例

３．研究の方法

(1)【平成 27 年度】

① 複数の要素を含んでいるデータベースを非巡回有向グラフに変換する手法の開発

既に文字列集合や組合せ集合に関しては系列二分決定グラフやゼロサプレス型二分決定グラフといったデータ構造があるので，XML などの木構造の集合を表現するデータ構造を研究していく．さらに，既存のデータ構造に対しても，より高速にグラフ表現へと変換する手法を開発する．

② 非巡回有向グラフを簡潔データ構造の組

合せで表現する方法の改良
研究代表者が 2014 年に提案した密集ゼロサプレス型二分決定グラフ(DenseZDD)をさらに改良し，通常のゼロサプレス型二分決定グラフと並行して扱うハイブリッド手法を実現する．研究代表者は，簡潔データ構造の権威である定兼邦彦教授の研究室に所属しており，高度なディスカッションを日常的に行うことのできる環境にある．氏と，氏の指導する学生とともにこの問題に当っていく．
③ 決められた節点数の非巡回有向グラフを一様ランダムに生成する手法の開発
非巡回有向グラフを対象とするアルゴリズムの開発において難しい点は実験データの確保である．こういった類のアルゴリズムの計算量はグラフの節点数に依存することが多いが，実データから望ましいサイズの非巡回有向グラフが得られるとは限らないためである．そこで与えられた節点数をもつグラフを一様ランダムに生成することができれば，アルゴリズムの性能実験も簡単かつスムーズに実施することができる．もしこれがうまくいかなかった場合は，逆に，n 変数論理関数を一様ランダムに生成した際にそれを表す二分決定グラフ，一種の非巡回有向グラフ，の平均サイズの解析を行うことで求めるサイズのグラフの作り方を考えていくこととする．
(2) 【平成 28 年度】
① 高速な検索を可能にする部分文字列索引の構築アルゴリズムを非巡回有向グラフ入力へ拡張
非巡回有向グラフによって圧縮して表現されている複数の文字列に含まれる全ての部分文字列を格納する部分文字列索引を構築するアルゴリズムを研究する．これは独力で一からつくり上げるものではなく，単一の文字列に対して考案された同目的のアルゴリズムで優れたものが多数存在するので，それらの技法を参考にしつつグラフへと拡張することによって実現する．
② 複数の文字列が非巡回有向グラフで与えられた時の近似文字列照合アルゴリズムの研究
複数の文字列が与えられた時に，それらの中から似ているもの，つまりできるだけ少ない回数の編集で同じ文字列になるもの，を発見したい場合がある．既存手法では，二つの文字列のペアを全て試して確認する手法や，全体をいっぺんに処理できる代わりにおおよそ正しい答えしか得られない手法などがあるが，本研究では非巡回有向グラフで文字列集合をまとめて処理しつつ正確な解を求める手法の開発を目指す．また，二つの文字列同士での近似文字列照合に関する慣例として，同じスコアの近似文字列が複数見つかった際には一つだけを選び出して計算を続け残りは捨ててしまう．しかし，選んだ一つが後続する計算において真に最適かどうかは保証されず，もしかしたら捨ててしまったものの中により良いものがあった可能性もある．そこで，本研究では発見した近似文字列を全て非巡回有向グラフの中に格納して続く計算もそのグラフ上で行うことで最終的な最適解を見落とすことなく求める方法を提案する．一般に，複数文字列をグラフで表現する際には，それらが互いに似ているほど圧縮して表すことができる．見つかった近似文字列同士も互いに類似していることが期待されるため本手法との相性は非常に良い．さらに，これは文字列だけではなく最適解を複数求めた後に一つだけを取り出してそれ以外は無視している他の既存手法にも適用可能な考え方で，文字列で実現した後はさらなる拡張を視野に入れている．
４．研究成果
(1) 【平成 27 年度】
① 複数の要素を含んでいるデータベースを非巡回有向グラフに変換する手法の開発について
既に文字列集合や組合せ集合に関しては系列二分決定グラフやゼロサプレス型二分決定グラフといったデータ構造があるので，その他の離散構造を表現するデータ構造を模索する.二分木の集合を表すグラフを開発している．
② 非巡回有向グラフを簡潔データ構造の組合せで表現する方法の改良について
所属研究室の教授と氏の指導していた学生とともに,筆者が 2014 年に提案した密集ゼロサプレス型二分決定グラフ(DenseZDD)をさらに改良したデータ構造の研究を行った．これは通常のゼロサプレス型二分決定グラフと完結データ構造によって圧縮して表現したものを並行して扱うハイブリッド手法であり,この結果はその学生によって国際学会 SEA 2016 で発表される予定である．
③ 決められた節点数の非巡回有向グラフを一様ランダムに生成する手法の開発について
非巡回有向グラフを対象とするアルゴリズムの開発において悩ましい点は実験データの確保である.そこで与えられた節点数をもつグラフを一様ランダムに生成することができれば,アルゴリズムの性能実験も簡単かつスムーズに実施することができる.二分決定グラフは多分木のペアによって表現できるため,そのような多分木対が有効である条件を調査し,そのランダム生成手法を考察した．
(2) 【平成 28 年度】
① 高速な検索を可能にする部分文字列索引の構築アルゴリズムを非巡回有向グラフ入力へ拡張について:部分文字列索引の構築だけではなく,より多種多様な非巡回有向グラフに対するアルゴリズムを数十種類提案し日本応用数理学会 2016 年度年会で発表した．ただ,計算量の解析は困難で,現在のところ一部のアルゴリズムに関してしかできていない.また,部分文字列索引の構築は一層の

効率化を実現できる感触があるので引き続き取り組んでいる.
② 複数の文字列が非巡回有向グラフで与えられた時の近似文字列照合アルゴリズムの研究について
近似文字列照合へのアプローチの一つとして,複数の文字列に対する中央値にあたるような文字列を近似的に発見するアルゴリズムを開発し国際学会 AAAC 2016 で発表した.
③ 開発したアルゴリズムを大規模実装し公開について
プログラムは実装したものの,それを整備し公開できる質に向上させるまでには到らなかった.
④ 非巡回有向グラフを簡潔データ構造の組合せで表現する方法の改良について
所属研究室の教授と氏の指導していた学生と研究を行いグラフが更新される場合にも対応できるようになった.この結果はその学生によって国際学会 SEA 2016 で発表された.さらに効率良く更新できるようなアルゴリズムの開発を進めている.

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）

〔学会発表〕（計 3 件）
① Shuhei Denzumi, Finding Approximate Median Strings Using Directed Acyclic Graphs, The 9th Annual Meeting of the Asian Association for Algorithms and Computation (AAAC 2016), Liang Kuo Shu International Conference Hall, College of Social Sciences, National Taiwan University (NTU), Taipei, Taiwan, 2016.
② Taito Lee, Shuhei Denzumi, Kunihiko Sadakane, Engineering Hybrid DenseZDDs, 15th International Symposium on Experimental Algorithms (SEA2016), St. Petersburg, Russia, 2016.
③ 伝住 周平, 系列二分決定グラフを用いた文字列集合演算, 日本応用数理学会 2016 年度 年会, 福岡県北九州市小倉北九州国際会議場, 2016.

６．研究組織
(1)研究代表者
伝住　周平（DENZUMI, Shuhei）
東京大学・大学院情報理工学系研究科・助教
研究者番号：90755729