様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成 24 年 5 月 1 日現在

機関番号：１３６０１
研究種目：基盤研究(C)
研究期間：2009～2011
課題番号：21500010
研究課題名（和文）文字列上のビット並列法を利用した高速木パターン照合アルゴリズムの開発

研究課題名（英文）Design of fast tree pattern matching algorithms using bit-parallelism on strings

研究代表者
山本　博章（YAMAMOTO HIROAKI）
信州大学・工学部・教授
研究者番号：10182643

研究成果の概要（和文）： 本研究課題では、木パターン照合問題を解くための効率的なアルゴリズムついて研究する。木パターン照合問題とは、各頂点が記号によってラベル付けされた２つのラベル付き木、すなわち、パターン木 P とデータ木 T が与えられたとき、T の中で P と一致する部分をすべて見つける問題である。本研究では、この問題に対し、文字列照合問題で開発されたビット並列法を木パターン照合問題に応用することにより、効率的なビット並列型木パターン照合アルゴリズムを開発した。

研究成果の概要（英文）： In this research, the following tree pattern matching problem is considered: Given two unordered labeled trees P and T, the problem is to find out all occurrences of P in T. Here P and T are called a pattern tree and a data tree, respectively. We developed efficient algorithms for the tree pattern matching problem by taking advantage of bit-parallelism on a string matching problem.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	1,300,000	390,000	1,690,000
2010 年度	1,300,000	390,000	1,690,000
2011 年度	800,000	240,000	1,040,000
年度			
年度			
総　計	3,400,000	1,020,000	4,420,000

研究分野：コンピュータサイエンス
科研費の分科・細目：情報学・情報学基礎
キーワード：アルゴリズム，パターン照合，木パターン，ビット並列，検索

１．研究開始当初の背景

木パターン照合問題は、バイオインフォマティクス、言語の処理系など情報科学の多くの分野で現れるため、それを解くための効率的なアルゴリズムの研究は、従来から活発に行なわれてきた。最近、XML と呼ばれるデータ形式が、データの保存・交換のための共通フォーマットとして広く認識されるようになり、XML データが急増している。XML データは木構造として表現できるため、ここでも木パターン照合アルゴリズムが重要な役割を演じている。このように、木パターン照合問題は今でも重要な研究課題となっている。木パターン照合問題は、大きく順序木、無順序木に分けることができる。以下で、それぞれに対する研究動向と本研究との関連について述べる。

(1) 順序木：従来の研究では、特に、親子関係及び兄弟の位置が完全一致するコンパクト照合条件と呼ばれる条件のもとでの効率的なアルゴリズムの開発に向けた研究が行われてきた。これらの研究の焦点は、パターン木の頂点数とデータ木の頂点数の積からなる計算時間の改良である。本研究は無順序木を対象とするが、そのアルゴリズムはコンパクト照合条件に対応した順序木の検索にも応用できる。このとき、パターン木の頂点数が少なければ、データ木の頂点数に比例した時間で動作し、従来のアルゴリズムより速い。

(2) 無順序木：木の類似度を求める問題、木の包含関係を求める問題について NP 完全性が示されている。特に、木の包含関係の問題は、後で述べる先祖子孫関係照合条件のもとでの木パターン照合問題と深い関係がある。最近、XML データの検索など無順序木が活用される場面も多く、無順序木に対する効率的な木パターン照合アルゴリズムが求められている。木に制限を課すことによって効率的なアルゴリズムを開発しようとする研究はあるが、順序木に比べ、効率的なアルゴリズムの開発が遅れている。本研究は、無順序木に対する効率的なアルゴリズムを開発する。

２．研究の目的

本研究課題では、親子関係照合条件と先祖子孫関係照合条件の２つの照合条件を導入し、これらの照合条件のもとでの無順序木に対する木パターン照合問題を考える。そのとき、本研究の目的は、文字列照合問題で開発されたビット並列法を応用し、次の点を明らかにすることである。

(1) 親子関係照合条件に対応したビット並列型高速木パターン照合アルゴリズムの開発
(2) 先祖子孫関係照合条件に向けたビット並列型高速木パターン照合アルゴリズムの開発
(3) 開発したアルゴリズムの実験的評価と XML データを用いた有効性の検証

３．研究の方法

まず、親子関係照合条件と先祖子孫関係照合条件を以下のように定義する。ここで、P をパターン木、T をデータ木とする。以下の定義は、T の中に P が出現する条件を述べている。

親子関係照合条件：(1) P の頂点から T の頂点への１対１対応が存在する、(2) P の任意の頂点に対し、同じラベルをもつ T のある頂点が対応する、(3) P の頂点の親子関係が保存される。すなわち、P の頂点 u が頂点 v の親ならば、T において u に対応する頂点は v に対応する頂点の親である。これら３つの条件を満たすものが親子照合条件である。

先祖子孫関係照合条件：上記の親子照合条件のうち(1)及び(2)の条件は同じであるが、条件(3)は次のようになる： P の頂点の先祖・子孫関係が保存される。すなわち、P の頂点 u が頂点 v の先祖ならば、T において u に対応する頂点は v に対応する頂点の先祖である。

本研究では、まず、親子関係照合条件のもとでのアルゴリズムを開発し、次にそれを先祖子孫関係へと拡張する。

３．１ 親子関係照合条件に対する木パターン照合アルゴリズムの開発

アルゴリズムの設計に当っては、文字列のパターン照合で開発された Shift-OR 法と呼ばれるビット並列法を利用する。このため、パターン木を文字列の形に分割する必要がある。本研究では、パターン木を根から葉への各パスで分割する「パス分割」と呼ばれる分割法を取る。まず初めに、本研究で導入する２つの重要な概念、「擬似木パターン照合条件」と「ラベルの再帰度」について説明する。

・擬似木パターン照合条件：これは、親子関係照合条件おける１対１対応を多対１対応に変えることによって定義される。このように条件を少し緩めるだけで効率的なビット並列型のアルゴリズムを設計することができる。
・ラベルの再帰度：ラベル付き木に対し、根から葉へ向かうパス上に出現する同じラベルの出現回数の最大値によって定義される。ラベルの再帰度はビット並列化の並列度と密接に関係し、本研究では、再帰度が小さいほど速いアルゴリズムを設計する。なお、アルゴリズムの計算時間は、パターン木のラベルの再帰度にのみ関係し、データ木の再帰度には関係しないという特徴を持つ。

以上の準備のもと、ビット並列型木パターン照合アルゴリズムを次の(1)及び(2)の 2 段階によって構成する。

(1) 擬似木パターン照合条件に対するアルゴリズムの開発：次の 2 ステップによる。
① パス分割によるパスの生成と Shift-OR 法のためのマスクパターンの作成
② Shift-OR 法を利用した上昇型木パターン照合アルゴリズムの設計： データ木上で葉から根に向かって照合操作を行

うアルゴリズムを設計する。パターン木を 分割した各パスをワードに収め、Shift-OR 法を利用した照合を行う。このとき、ビット並列化のため、

- 各パスそれぞれを１ワードに収める、
- すべてのパスをまとめて１ワードに収める、

の２種類の手法について検討する。また、パスのチェックだけでは兄弟間の関係がチェックできないため、兄弟間で同期を取るためのビット並列法も導入する。

(2) 擬似木パターンから親子関係照合条件を満足する木パターンの抽出

① パターン木の兄弟間のラベルがすべて違う場合：擬似木パターン照合条件の多対１対応はすべて１対１対応になるため、上記(1)の結果が親子関係照合条件の結果となる。

② 同じラベルがある場合： 親子関係照合条件の１対１対応を満足するかチェックする必要がある。このために、2 部グラフの最大マッチングを求めるアルゴリズムを利用する。

３．２　先祖子孫関係照合条件に対する木パターン照合アルゴリズムの開発

親子関係照合条件に対する照合アルゴリズムを拡張することによって、先祖子孫関係照合条件に対するアルゴリズムの開発を行なう。拡張のキーポイントは次の 2 点になる。

(1) 擬似木パターン照合条件の導入：多対１対応を用いた先祖子孫関係照合条件に 対する擬似木パターン照合条件を定義する。

(2) 先祖子孫枝の導入による照合アルゴリズムの拡張： 任意の文字にマッチするドントケア記号というものがある。ここでは、これに対応した先祖子孫枝というものを導入する。すなわち、パスの中に親子関係だけでなく、先祖子孫関係を示す枝を導入することにより、先祖・子孫関係をチェックするようにアルゴリズムを改良する。Shift-OR 法においてもドントケア記号を考慮した改良が行われており、この考えを利用することにより、アルゴリズムを開発する。

４．研究成果

本研究課題では、木パターン照合問題に対し、以下のように効率的なビット並列型木パターン照合アルゴリズムを開発した。

(1) 擬似木パターン照合を求めるためのビット並列アルゴリズムの開発：親子関係に基づいた擬似木パターン照合問題を定義し、この問題に対し効率的なビット並列アルゴリズムを与えた。擬似木パターンはXML検索において重要な役割を演じている。アルゴリズムの基本的な考えは、パターン木をパスパターンと呼ばれる記号列に分割し、各パスパターンをビット列としてコンピュータワードの中に埋め込むことによって、ビット並列化を実現するものである。埋め込む手法により、以下の２つのビット並列アルゴリズムを与えた。本アルゴリズムの計算時間は、パターン木の再帰度によって特徴づけができ、再帰度が小さいほど高速に動作する。さらに、どちらも、パターン木が小さい場合、データ木のサイズに比例した時間で動作する。

① パスパターンを直列的にコンピュータワードに埋め込むことにより高速に動作するビット並列アルゴリズム

② 各パスパターンをビット単位で相互に互い違いに並べて埋め込むことにより高速に動作するビット並列アルゴリズム

(2) 擬似木パターン照合から木パターン照合を求めるためのアルゴリズムの開発： 2部グラフの最大マッチングを求めるアルゴリズムを応用することにより、擬似木パターン照合から木パターン照合を求めるためのアルゴリズムを与えた。

次に上記の結果を先祖子孫関係へ拡張するため、拡張擬似木パターン照合問題を定義し、以下のように効率的なビット並列型拡張擬似木パターン照合アルゴリズムを開発した。

(1) 拡張擬似木パターン照合問題の導入： 親子関係に基づいた擬似木パターン照合問題を先祖子孫関係に拡張した拡張擬似木パターン問題を導入した。具体的には、パターン木を記述するにあたって、通常の親子関係を示す枝に加え、先祖子孫関係を示す枝（先祖子孫枝）を導入することにより擬似木パターン照合問題を拡張した。拡張擬似木パターン照合問題は、XML検索におけるXPathの照合問題と密接に関係しており、重要な問題となっている。

(2)拡張擬似木パターン照合問題に対するビット並列アルゴリズムの開発： 擬似木パターン照合問題と同様の計算量で動作するビット並列型アルゴリズムを開発した。具体的には、擬似木パターン照合問題と同様に、パターン木をパスパターンに変換する方法により以下の①及び②の2種類のアルゴリズムを開発した。どちらもコンピュータワードに収ま

るパターン木に対し、高速に動作する。

① パターン木をパスに分割し、各パスに対するビットパターンを順番に直列的に並べてコンピュータワードに埋め込むことにより高速に動作するビット並列アルゴリズムを開発した。

② パターン木をパスに分割し、各パスに対するビットパターンをビット単位で互い並列に並べてコンピュータワードに埋め込むことにより高速に動作するビット並列アルゴリズムを開発した。

XMLへの応用を考えた場合、XMLに対応する木構造は記号列でラベル付けされている。しかし、上記で開発したアルゴリズムは記号でラベル付けされた木を考えている。そこで、拡張擬似木パターン照合問題を、木のノードが記号列でラベル付けされた問題に拡張し、それに対するアルゴリズムを開発した。具体的には以下の2点に関する結果を得た。

(1) 記号列をラベルに持つ拡張擬似木パターン照合問題に対するアルゴリズムの開発：
XMLデータは、要素名をラベルとする木構造で表すことができる。すでに開発したアルゴリズムは記号をラベルとして持つ木構造に対するものであったため、それらを記号列をラベルとして持つ木の照合問題へと拡張した。アルゴリズムは、まず、パターン木のラベルから、そのラベルを識別するためのオートマトンを作成する。そのあと、データ木からパターン木に一致する部分木を探すとき、データ木のノードのラベルについてはこのオートマトンを用いて識別する。また、パターン木の検索に対しては、今までに開発したビット並列法を使って照合を行う。

(2) 記号列照合のためのコンパクトな決定性有限オートマトンの開発： 記号列をもつ拡張擬似木パターン照合問題では、ラベルとなる記号列の高速な認識が必要となる。決定性有限オートマトンは効率的なパターン照合を実現するが、そのサイズが大きくなることが問題となる。そこで、決定性有限オートマトンのコンパクトな実現方法を開発した。具体的には、双対型positionオートマトンと呼ばれるモデルを利用し、決定性有限オートマトンをコンパクトなビット列で表す方法を示した。 今回の手法は、不要な状態を取り除き、できるだけコンパクトなものを実現した。また、正規表現と呼ばれる一般的な文字列パターンに適用することができる。

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）

〔雑誌論文〕（計１件）

① H. Yamamoto and D. Takenouchi: Bit-parallel Tree Pattern Matching Algorithms for Unordered Labeled Trees; Proc. of WADS 2009, LNCS 5664, 554-565, 2009, 査読有
http://hdl.handle.net/10091/13658

〔学会発表〕（計５件）

① 山本博章, 中村彰吾：双対型 position オートマトンを用いたコンパクトな DFA 表現；電子情報通信学会技術報告, COMP2011-36, 1-7, 2011/12/16, 名古屋
② 山本博章, 宮嵜敬：記号列のラベルをもつ拡張擬似木パターンマッチング；電子情報通信学会技術報告, COMP2011-25, 53-60, 2011/9/6, 函館
③ 坂田俊則, 山本博章：正規表現からコンパクトな有限オートマトンを構成するための一手法；電子情報通信学会技術報告, COMP2010, 2010/9/29, 長岡
④ 山本博章, 宮嵜敬：拡張擬似木パターンマッチング問題に対するビット並列アルゴリズム；情報処理学会研究会報告, AL-2010-131-4, , 2010/9/22, 函館
⑤ 山本博章：無順序木パターン照合に対するビット並列アルゴリズム, 冬の LA シンポジウム（京都）, (17-1)-(17-10), 2010/2/2

６．研究組織

(1)研究代表者
山本　博章（YAMAMOTO HIROAKI）
信州大学・工学部・教授
研究者番号：10182643

(3)連携研究者
岡本　正行（OKAMOTO MASAYUKI）
信州大学・工学部・教授
研究者番号：50109196

白井　啓一郎（SHIRAI KEIICHIRO）
信州大学・工学部・助教
研究者番号：00447723

(4)研究協力者
宮嵜　敬（MIYAZAKI TAKASHI）
長野工業高等専門学校・教授
研究者番号：10141889