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２５年　６月　２５日現在

機関番号：３４３１０
研究種目：若手研究（B）
研究期間：2009～2012
課題番号：２１７００１７８
研究課題名（和文）　木構造データに対する汎用性の高い類似度計算技術の開発

研究課題名（英文）　Development of versatile techniques for computing similarity between tree-structured data

研究代表者
深川　大路（FUKAGAWA DAIJI）
同志社大学・文化情報学部・助教
研究者番号：10442518

研究成果の概要（和文）：本研究では，文字列データと比較してより複雑な木構造データに対して，距離の計算や近似マッチングを行うための汎用的な方法を開発することを目的として，順序木および無順序木に対する高速近似アルゴリズム，動的計画法を用いた厳密アルゴリズム，および，最大クリーク問題への帰着と高速な専用ソルバーを利用したプログラム，木構造のマッチングに応用可能な確率的生成モデルの提案および学習アルゴリズムなどの開発および理論的解析を行った．

研究成果の概要（英文）：In this project, we developed several methods which computes distance between tree structured data and gives an approximate matching between them. As a result, we obtained fast approximate algorithms which computes edit distance between ordered/unordered trees, fixed parameter algorithm to compute the proper tree edit distance, fast exact algorithm to compute an optimal tree matching via maximum clique problem, a probabilistic model to measure similarity of trees and a learning algorithm for the model.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００９年度	700,000	210,000	910,000
２０１０年度	600,000	180,000	780,000
２０１１年度	500,000	150,000	650,000
２０１２年度	500,000	150,000	650,000
総　計	2,300,000	690,000	2,990,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：木構造，類似度，近似パターン照合，アルゴリズム

１．研究開始当初の背景

生物情報学や自然言語処理，画像解析をはじめとするさまざまな分野で知識発見や類似度の計算手法が開発されてきた．なかでも，編集距離の概念を用いた手法は，パターン照合や類似度計算のための標準的な手法として広く受け入れられており，様々な応用分野で実際に使われている．一方で，構造を持つデータに対するパターン照合については様々な技術が枠組みとして提案されているものの，実用レベルにおいて汎用的あるいは標準的な方法は知られていない．ある応用においてどの技術を選択し組み合わせるかについての判断は専門家にとっても容易ではなく，アドホックな判断が必要である．

研究代表者は，木の編集距離を計算する問題に着目し，これまでに厳密アルゴリズムおよび近似アルゴリズムとして様々な手法を

提案してきた．順序付けが一意に定まらない無順序木に対しては，木の編集距離計算を含む多くの問題が近似困難であると証明されている [Bille 2005]．この問題に対して，研究代表者等は近似アルゴリズムの開発，最大共通部分木問題に対する O(1.5h) 近似アルゴリズムの開発 (h は木の高さを表すパラメータ)，編集距離の特殊例である bottom-up mapping を計算するための厳密アルゴリズムの開発，などを行なってきた．

木構造の類似度計算は，データベースにおける検索を含む，幅広い応用分野を持っている．木の編集距離では多くの場合，単位コストによる距離計算が用いられてきた．一方で文字列の場合，例えば生物情報配列の近似パターン照合においては，単位コストではなく，PAM や BLOSUM 等のスコア行列が広く利用されている．これらのスコア行列の価値は経験的に十分評価されている．木の編集距離計算においては同様のスコア行列やその獲得方法は確立しておらず，上記のように多くの場合は単位コストが用いられる．研究代表者等は，木の編集距離の一種である tree alignment に基づいて確率的類似度モデルを構築し，そのモデルのパラメータを学習するための手法を提案した．この手法は，データベース間のスキーママッチングを行う際，少ない事例から最適なパラメータを学習する等の目的に適用可能である．しかし，上記の確率モデルは，XML 等の豊富な非構造情報を含むような場合には直接適用する事は困難である．また，学習アルゴリズムにおいても，既にいくつかの代表的な機械学習アルゴリズムを適用可能であることは分かっているものの，木構造データに対する適用事例は少なく，選択肢は乏しい．

このような状況において，汎用的な木構造の類似度計算，距離（非類似度）計算，マッチングなどのアルゴリズムを開発することへの要求および重要性は高いと考えられる．

２．研究の目的

本研究の目的は，「木構造に対する近似パターン照合アルゴリズム」および「木構造の確率的類似度モデルに対するパラメータ学習」の二つに大別される．それぞれ詳細を以下に示す．

(1) 木構造に対する近似パターン照合アルゴリズム

研究代表者等がこれまでに提案した手法は，木の近似パターン照合および類似度計算問題における代表的な部分問題を適用対象とするものの，標準的な場合に適切なアルゴリズムを選択するといった応用との結びつきには考察の余地が残されていると考える．例えば，XML 等の半構造データは，高さが制限された構造を持ち，かつラベルに関しても内部ノードの持つ情報が疎である一方で葉ノードに集中して情報が含まれる等，極めて特殊な性質を持つ．

本課題においては，データの分野や性質に応じて性能の高いアルゴリズムを選択する事は応用において需要の高い課題であると捉えて研究を行う．既存研究において提案されたアルゴリズムの性能を事例に基づいて再評価するとともに，その結果によって有効性が予想される新たなアルゴリズムを開発する．

(2) 木構造の確率的類似度モデルに対するパラメータ学習

XML のような半構造データは，葉ノードに多くの非構造化情報が集中するような特殊な性質を持っているため，確率モデルにもこのような性質を反映させる事により，より良い類似度としての尺度が得られる可能性がある．研究代表者等が提案した tree alignment に基づく確率的生成モデルは葉ノードにおける上記の特殊な性質を用いていないが，生成モデルに組み込むことは可能である．

３．研究の方法

これまで提案したアルゴリズムについて，計算機実験をもとにそれら性能を精緻に再検討したうえで，改良を試みる．同時に，実世界に存在する木構造データの性質についても検討し，アルゴリズムの改良に対して効果の高い性質を見つけ出し，それに着目して研究を進める．例えば，書誌情報をはじめとして現在普及している XML データ，および公開データセットの多くは，階層の比較的浅い特殊な木構造を有する事が申請者等によって指摘されている．したがって，そのような特殊な場合において特に高い性能を持つ類似検索アルゴリズムの開発は，本計画の目的において非常に重要な課題であると考えられる．

４．研究成果

(1) 順序木のマッチングに対する近似アルゴリズムの開発 [論文⑤]

編集距離の計算量が比較的少ない順序木の場合においてさえ，知られている限りにおいて最良の時間計算量 $O(n^3)$ は十分に実用的な速度を持つとはいえない．さらに，この時間計算量は一般的な仮定のもとにおいて下限と一致することが示されており，改善は容易ではない．このため，時間計算量の改善を目的とする研究においては近似アルゴリズムの開発も行われてきた．著者らは，modified Euler string への変換を利用した近似アルゴリズムを提案し，そのアルゴリズムの理論的性能の解析を行った．その結果と

して，アルゴリズムが近似度 O(n^0.75) および時間計算量 O(n^2) を持つことを証明した．多くの既存アルゴリズムが存在するなかで，著者らの提案するアルゴリズムは，以下の特徴において優れていると考えられる：a) 基本的なアイデアが単純であり，理解しやすく，実用や解析および改良に向いている；b) 理論的な近似性能保証を持つ；c) 計算時間は変換後の編集距離の計算時間に単純に依存しており，問題変換そのものの時間計算量は O(n) であり高速に動作すると考えられる．

(2) 無順序木のマッチングに対する高速な近似アルゴリズムの開発［雑誌論文①，学会発表⑦⑨⑩］

無順序木の編集距離については順序木とは異なり，多項式時間アルゴリズムの開発自体が困難である．また，その性質から，順序木で用いた近似アルゴリズムのアイデアをそのまま適用することもできない．したがって，高速な近似アルゴリズムの開発が有効であると考えて開発に取り組んだ．その結果として，与えられた木の高さを h としたときに近似度 2h+2 を保証するアルゴリズムの開発に成功した．著者らが提案したこの近似アルゴリズムは，無順序木の編集距離を高次元特徴空間上の L1 距離に埋め込むことによって高速計算を可能にしたという点が主なアイデアであり，その特徴ベクトルは，部分木の頻度を用いた．このようにして得られる距離尺度は bottom-up distance として既に知られていたものと等価であると考えられるものの，近似性能に関する理論的保証については考慮されていなかった．これを明らかにしたのが本研究の成果である．

(3) 無順序木に対する高速な固定パラメータ・アルゴリズムの開発［雑誌論文①④，学会発表①②⑧］

固定パラメータ・アルゴリズムは，解の厳密性を保証するアルゴリズムの一種である．無順序木の近似アルゴリズムは，高速である一方，十分な近似性能が得られない可能性を含んでいる．大規模データに対する絞込み等に代表されるいくつかの場面においては，必ずしも厳密解を求める必要はなく，近似アルゴリズムが十分に有用であると考えられるが，その一方で，厳密解の保証が要求される場面においては，他の方策を検討することになる．本研究においては，特定のパラメータを固定した場合，具体的には，編集距離の値の上限を k，木の頂点数の上限を n とした場合において，O(2.62^k・poly(n)) 時間で編集距離の厳密な値を計算できるアルゴリズムを開発した．このアルゴリズムは動的計画法を用いて最大共通部分木を計算するものであり，その最悪時計算量は Fibonacci 数列を用いた解析によって得られた．その後，改良によって O(1.26^2n) 時間，O((1+ε)^2n) 時間，O(2^2b・poly(n)) 時間など，いくつかのパラメータについて異なる計算量をもつアルゴリズム等をそれぞれ得た．

(4) 最大クリーク問題への帰着を利用するアルゴリズム［雑誌論文②③，学会発表③④⑥］

固定パラメータ・アルゴリズムと同じく厳密アルゴリズムの枠組みにおいて，実用的な性能を持つソフトウェアの開発を目的として，研究を行った．本研究では，木編集距離問題を最大クリーク問題に帰着させ，最大クリーク問題に対する高速ソルバーを利用して元の木編集距離問題を解くという手法を開発した．最大クリーク問題は，NP 困難であることが証明されている問題のひとつである．最大クリーク問題のソルバーとして，現在最速として知られているもののひとつを，ソルバー開発グループから提供を受けて利用した．提案手法の利点は，実用的な時間内で実問題を扱える点と，厳密解が保証されている点，さらに，編集距離のマッチング・コストを単位コストに限定しない点である．本研究の成果として得られたものは次の通りである．既存の高速な探索的アルゴリズムと比較して劣らない程度に高速で，かつ厳密解が保証される計算機プログラムが得られた．また，糖鎖データ等に対する計算機実験を行い，その性能を評価した．さらに，その提案手法の改良に取り組み，ヒューリスティクスを工夫することにより高速化を実現し，その効果を，実データに対する計算機実験によって検証した．

(5) 木構造の確率的類似度モデルに対するパラメータ学習［学会発表⑤］

テキストデータに対して広く利用されている確率的生成モデルのベイズ学習を，木構造のような階層を持つデータに対して適用できるよう拡張することを試みた．木構造に対しては，既に，確率的生成モデルおよび最尤推定によるモデル学習アルゴリズムを開発していた．しかし，木構造はテキスト情報以上にパラメータの個数も多いため，過学習を避けてモデルの学習精度をより向上するためにはベイズ学習を用いるなどの工夫が必要であると考えられる．各パラメータが事前分布を持てるようにモデルを改良するとともに，変分ベイズ法を利用して学習アルゴリズムを開発した．

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５件）

① Tatsuya Akutsu, Daiji Fukagawa, Magnús M. Halldórsson, Atsuhiro Takasu, Keisuke Tanaka. Approximation and parameterized algorithms for common subtrees and edit distance between unordered trees. Theoretical Computer Science, 470:10-22, 2013（査読有）. DOI:10.1016/j.tcs.2012.11.017

② Tomoya Mori, Takeyuki Tamura, Daiji Fukagawa, Atsuhiro Takasu, Etsuji Tomita, Tatsuya Akutsu. A Clique-Based Method Using Dynamic Programming for Computing Edit Distance Between Unordered Trees. Journal of Computational Biology, 19(10):1089-1104, 2012（査読有）. DOI:10.1089/cmb.2012.0133

③ Daiji Fukagawa, Takeyuki Tamura, Atsuhiro Takasu, Etsuji Tomita, Tatsuya Akutsu. A clique-based method for the edit distance between unordered trees and its application to analysis of glycan structures. BMC Bioinformatics 12(S-1): S13, 2011（査読有）. DOI:10.1186/1471-2105-12-S1-S13

④ Tatsuya Akutsu, Daiji Fukagawa, Atsuhiro Takasu, Takeyuki Tamura. Exact algorithms for computing the tree edit distance between unordered trees. Theoretical Computer Science, 412(4-5): 352-364, 2011（査読有）. DOI:10.1016/j.tcs.2010.10.002

⑤ Tatsuya Akutsu, Daiji Fukagawa, Atsuhiro Takasu. Approximating Tree Edit Distance through String Edit Distance. Algorithmica 57(2): 325-348, 2010（査読有）. DOI:10.1007/s00453-008-9213-z

〔学会発表〕（計１０件）

① Tatsuya Akutsu, Takeyuki Tamura, Daiji Fukagawa, Atsuhiro Takasu. Efficient Exponential Time Algorithms for Edit Distance between Unordered Trees. Proc. 23rd Annual Symposium on Combinatorial Pattern Matching (CPM 2012), Helsinki, Finland, July 3-5, 2012.

② 阿久津達也，田村武幸，深川大路，高須淳宏．無順序木の編集距離の指数時間厳密アルゴリズム．電子情報通信学会コンピュテーション研究会，2012 年 06 月 21 日，北海道大学.

③ Tomoya Mori, Takeyuki Tamura, Daiji Fukagawa, Atsuhiro Takasu, Etsuji Tomita, Tatsuya Akutsu. An Improved Clique-Based Method for Computing Edit Distance between Rooted Unordered Trees. 情報処理学会第 26 回バイオ情報学研究発表会，2011 年 9 月 13 日，神戸大学.

④ Tatsuya Akutsu, Tomoya Mori, Takeyuki Tamura, Daiji Fukagawa, Atsuhiro Takasu, Etsuji Tomita. An Improved Clique-Based Method for Computing Edit Distance between Unordered Trees and Its Application to Comparison of Glycan Structures. International Conference on Complex, Intelligent and Software Intensive Systems (CISIS 2011), June 30 - July 2, 2011, Korean Bible University (Seoul, Korea).

⑤ Atsuhiro Takasu, Daiji Fukagawa, Tatsuya Akutsu. A Variational Bayesian EM Algorithm for Tree Similarity. Proc. 20th International Conference on Pattern Recognition (ICPR 2010), Istanbul, Turkey, 23-26 August 2010.

⑥ 深川大路，田村武幸，高須淳宏．無順序木の編集距離を計算する実用的アルゴリズムについて，人工知能学会 第 78 回人工知能基本問題研究会．2010 年 8 月 1 日，兵庫県立大学(神戸キャンパス).

⑦ 深川大路，阿久津達也，高須淳宏，安達淳：高さ制約付き無順序木の高速類似検索アルゴリズムについて，情報処理学会第 72 回全国大会．2010 年 3 月 9 日，東京大学.

⑧ 阿久津達也，深川大路，高須淳宏，田村武幸．無順序木の編集距離計算のための厳密アルゴリズム．情報処理学会アルゴリズム研究会，2010 年 3 月 5 日．東芝科学館(東京).

⑨ Daiji Fukagawa, Tatsuya Akutsu, Atsuhiro Takasu. Constant Factor Approximation of Edit Distance of Bounded Height Unordered Trees. 16th International Symposium on String Processing and Information Retrieval (SPIRE 2009), pp.7-17, Saariselkä, Finland, August 25-27, 2009.

⑩ 深川大路，阿久津達也，高須淳宏．高さの制限された無順序木の編集距離問題に対する近似アルゴリズム．電子情報通信学会コンピュテーション研究会，2009 年 6 月 29 日,北海道大学.

６．研究組織

(1)研究代表者

深川　大路（FUKAGAWA DAIJI）

同志社大学・文化情報学部・助教

研究者番号：10442518