確率モデルに基づく木の類似度のパラメータ学習について

深川大路† 高須淳宏† 阿久津 達也‡

†国立情報学研究所 ‡京都大学化学研究所

1 はじめに

木の編集距離は，木を比較する際の距離尺度として広く用いられている．文字列の編集距離の場合と同様に，木の編集距離を計算するためには，各編集操作のコストを事前に与える必要があり，計算結果はコストの設定に大きく依存する．近年，木構造をともなうデータが様々な分野で用いられ，木に対するデータマイニングが注目されている．例えば，XML 文書の分類，Web からの情報抽出，XML ストリームに対するパターン抽出等である．

本研究では，木のアラインメントに基づく確率モデル vertex expansion HMM を提案し，正例集合からパラメータ学習するためのアルゴリズムを提案する．パラメータ学習には EM アルゴリズムを用いた．

2 木の編集距離

データマイニングにおいては，与えられたデータ同士の類似度を測ることがしばしば必要となる．木の類似度を測るための尺度として，木の編集距離が一般に用いられる．与えられた二つの木に対して編集距離を計算するためのアルゴリズムは古くから研究されており，多くのアルゴリズムが知られている [2, 3].

木の編集距離を計算するためには，文字列の編集距離を計算するに比べて多くの計算コストが必要である．そのため，近似的な編集距離を類似度をより効率的に計算するための手法が提案されている．

これらの近似アルゴリズムの計算効率は，“filter and refinement” におけるフィルターとして非常に有効であるが，refinement フェーズにおいては，より精度の高い類似度尺度を用いる必要がある．また，木の編集距離を求める際には，各編集操作のコストを事前に与える必要があり，そのコストが類似度として精度に影響する．データに関する知識を用いて人手によってコストを与えるには多くの労力を要するうえに，データの種類と量が増えるにしたがって現実的には計算機による計算で求める事は不可欠となる．本稿では，計算機によってこのようなパラメータ（編集操作のコスト）を求めるための手法を提案する．文字列の編集距離に関しては編集距離のコストを求める手法 [4] が知られており，この方法で求めたパラメータは情報統合タスクにおいて効果的であるという結果も得られている [1]. 本研究はこの手法を木に拡張するものである．

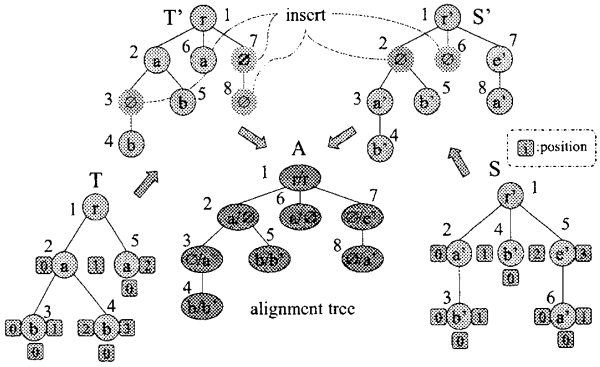

図 1: Example of tree alignment.

本稿では根付き順序木，すなわちある一つの頂点を根とし，各頂点の子供の間に全順序が与えられている場合のみを扱う．根付き順序木を単に木と呼ぶことにする．また，各頂点にはラベルが与えられているものとする．木の編集操作は置換・挿入・削除の三種類を考える．

3 木の確率的アラインメントモデル

木のアラインメント 木のアラインメント [3] は以下のように定義される．与えられた木の対 T, S に対して，それぞれの木に null ラベル ϕ を持つ頂点をいくつか任意の場所に挿入し，同じ構造を持つ木 T' と S' を作る．このようにしてできる木の対を T と S のアラインメントと呼ぶ．木のアラインメントにおいて対応する頂点のラベルを対にして得られる木をアラインメント木と呼ぶ．図 1 にアラインメント木の例を示す．

Learning Tree Similarity Based on a Statistical Model

Daiji Fukagawa†, Atsuhiro Takasu†, and Tatsuya Akutsu‡
†National Institute of Informatics, Tokyo 101-8430, Japan
‡Bioinformatics Center, Institute for Chemical Research, Kyoto University, Kyoto 6110011, Japan
{daiji, takasu}@nii.ac.jp takutsu@kuicr.kyoto-u.ac.jp

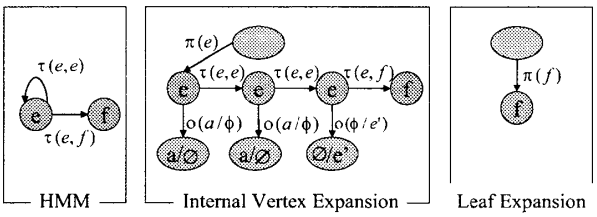

図 2: vertex expansion HMM の例

確率モデル 本研究では，アラインメント木の各頂点から部分木を生成する隠れマルコフモデル（HMM）を用いる．この HMM を *vertex expansion HMM* と呼ぶ．vertex expansion HMM は木の生成モデルである．生成の様子を図 2 に示す．

頂点 v に対して，v とその子（子の子やその子孫は除く）よりなる木が生成される確率を，頂点 v の展開確率とよび，$\Pr(v;\theta)$ とかく．ただし，θ は HMM のパラメータである．定義より，$\Pr(v;\theta)$ は外部の構造によらず決まる．このときアラインメント木 A が生成される確率は $\Pr(A;\theta) = \prod_{v\in V(A)} \Pr(v;\theta)$ とかける．これにより，木 T と S の類似度を周辺確率によって

$$\Pr(T,S;\theta) = \sum_{A} \Pr(A;\theta) = \sum_{A} \prod_{v\in V(A)} \Pr(v;\theta)$$

と定義する．ただし，右辺の和はペア (T,S) の間に可能な全てのアラインメント木について考える．

4 パラメータ学習

次に，学習データ集合からパラメータを学習するための手法について概要を述べる．学習データ集合 $\mathcal{D}$ は互いに類似度の高い木の対の集合である．パラメータ θ は，対数尤度 $\sum_{\langle T,S\rangle\in\mathcal{D}} \log \Pr(T,S;\theta)$ を最大にするように選べばよい．解析的には解くことが困難であるが，他の HMM と同様に EM アルゴリズムによる推定が可能である．

上の式において全ての組み合わせについて計算するためには指数オーダの計算を行うことになり実用上困難であるが，アルゴリズムを工夫することによって計算コストを抑えることが出来る．木 T と S のあらゆるアラインメント木を考えるとき，頂点の対 $u\in V(T)$, $v\in V(S)$ がアラインメント木 A に出現する頻度の期待値を $\mathcal{E}_{\mathrm{s}}(u,v)$ とおく．また，頂点 $u\in V(T)$ が削除され，木 S の位置 $v_{i:j}$ に対応するような場合が起こる頻度の期待値を $\mathcal{E}_{\mathrm{i}}(u,v_{i:j})$ とおく．同様に，木 T の位置 $u_{i:j}$ に頂点 $v\in V(S)$ が挿入される頻度の期待値を $\mathcal{E}_{\mathrm{d}}(u_{i:j},v)$ とおく．すると，EM-step の更新式はこれら三つの期待値によって表される．

パラメータ推定のためのステップは，以下の通りである:

- これら三種類の期待値を全ての場合について計算しておき（E-step），
- それをもとに更新式を計算する（M-step）．

三種類の期待値をすべての場合について計算するために，対となる頂点の子孫（内側）とそれ以外（外側）に分割し，それぞれの確率を計算する．

学習データにおける木の頂点数と次数（子の個数）がそれぞれ n, d 以下であるとする．計算すべき期待値の個数は $O(n^2d^4)$ 個である．これらは動的計画法により $O(n^2d^6)$ の計算時間によって求める事ができる．各パラメータの再推定をための計算時間量はこれより少ないオーダですむ．学習データの個数を m とすると，EM-step は $O(mn^2d^6)$ の計算時間で実行できる．領域量については全体で高々 $O(n^2d^4)$ 個の確率（または期待値）を保持すればよい．

5 まとめと今後の課題

本研究では，木の編集距離を類似度として使うためのパラメータを正例集合から学習するための確率モデルとアルゴリズムを提案した．提案手法により，様々な用途に適した木の類似度尺度となるモデルを効率的に学習できるため，データマイニングなどへの応用が考えられる．また，人工データに対していくつかの実験を行った（本稿では紙面の都合により省略）．

アラインメント木の生成には vertex expansion HMM と呼ぶ単純な確率的生成モデルを用いた．今後は，より複雑なモデルを考えることが考えられる．現在は，文字列をラベルとしてもつ場合への対応が困難である．また，順序木のみを扱っている．実際には無順序木として扱うべき場合も多いため，新たなモデルでは部分木の入れ換えを許すアルゴリズムを考えるのが今後の課題である．

参考文献

[1] M. Bilenko and R. J. Mooney. *KDD03*, 2003.

[2] P. Bille. A survey on tree edit distance and related problems. Theor. Comput. Sci., 2005.

[3] T. Jiang, et al. *TCS*, 14(3):137–148, 1995.

[4] E. S. Ristad and P. N. Yianilos. Learning string-edit distance. *IEEE Trans. on PAMI*, 1998.