D-004

部分構造の相対的な位置関係を考慮した順序あり木の類似検索

大高一輝 †　　古賀久志 †

1　はじめに

木はコンピュータサイエンスにおける基本的なデータ構造である．木は根の有無やノードの順序の有無によっていくつか種類があるが，本研究では根あり順序あり木を用いる．根あり順序あり木は RNA や XML を表現するのに用いられている．木の類似検索は，例えば構造の似た RNA や XML，タンパク質を探すといった応用で用いられる．木のような構造を持つデータ構造に対する類似検索では，

- データ間の類似度をどう定義するか
- 類似度をどう高速計算するか

が重要である．近年では計算量が大きい木編集距離の代わりに，木を部分構造集合に変換することで木間類似検索を集合間類似検索に帰着して解くアプローチが見られる．集合間類似度としては Jaccard 係数がよく使われている．具体例としては，Nikolaus ら [1] は順序あり木を pq-gram で表される部分構造の集合として表現した．Peisen ら [4] はさらに Min-Hash と Hadoop を用いて部分構造集合に対する類似検索を高速化した．

しかし上記のように， 木を部分構造の集合として表す手法では，木の局所情報のみに注目しており大域的な情報を捨ててしまっている．例えば，部分構造が木内のどこに存在するのかという情報は失われる．そこで，本研究では順序あり木を (部分構造のペア，その相対的な位置関係) の集合として表す手法を提案する．

第 2 節で従来手法とその問題点を述べる．次に第 3 節で提案手法を述べた後，第 4 節で評価実験を行い，第 5 節で結論を述べる．

2　従来手法: pq-gram を用いた木間類似度

木編集距離は根あり順序あり木に対する汎用的な類似度評価手法である．木 T_1，T_2 に対して T_1 と T_2 の編集距離とは，T_1 を T_2 に変換するために必要な最小の編集距離コストである．木を変換するための操作には主に挿入，削除，置換操作の 3 つが用いられており，各操作にはコストが定義されている．T_1 を T_2 に変換する編集操作列を S，その編集操作の合計コストを $\gamma(S)$ とすると，T_1 と T_2 の木編集距離は式 (1) のように定義される [2].

$$\delta(T_1, T_2) = \min\{\gamma(S) | S \text{ は } T_1 \text{ から } T_2 \text{ への木変換 }\} \tag{1}$$

木編集距離の計算量は木のノード数 n に対し

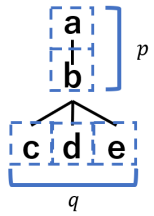

図 1　$p = 2$, $q = 3$ の場合の pq-gram

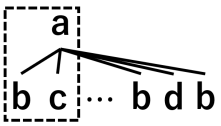

図 2　pq-gram の生成例 1

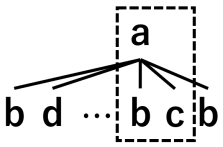

図 3　pq-gram の生成例 2

$O(n^3 \log n)$ と大きい．そこで木を部分構造の集合として表現し，集合間類似度を計算する手法が研究されてきた．Nikolaus ら [1] や Peisen ら [4] は根あり順序あり木 T に対して pq-gram と呼ばれる部分木 t を全て取り出し，pq-gram 集合 $S = \{t_1, t_2, \cdots, t_{|S|}\}$ として木を表現することで，その集合間類似度を木間類似度として高速に計算した．pq-gram は図 1 のような部分木のことで，p 個の非葉ノードと，その最下部に q 個の葉ノードを持つ．

pq-gram 集合 S_1，S_2 間の pq-gram 類似度 Sim は式 (2) のように Jaccard 係数で定義される．

$$\mathrm{Sim}(S_1, S_2) = \frac{|S_1 \bigcap S_2|}{|S_1 \bigcup S_2|} \tag{2}$$

加えて，LSH の一種である Min-Hash はハッシュ値の一致確率が Jaccard 係数と等しいため，Min-Hash を用いた類似度計算のさらなる高速化も可能である [3].

2.1　問題点

従来研究では，木を部分構造の集合として表現するが，その際に部分構造が木のどこに位置していたかという情報が失われている．

式 (2) で示される pq-gram 類似度 Sim は集合の要素が一致すれば類似度が高まるが，2 つの木が同一の pq-gram を持っていたとしても似ているとは限らない．

例えば $p = 1$, $q = 2$ の時，図 2 の点線で囲まれた pq-gram は木の左側で生成されているが，図 3 の点線で囲まれた pq-gram は右側で生成されている．これらの pq-gram は，形状は一致しているが生成された位置が大きく異なっているため，類似しているとは言えない．

† 電気通信大学大学院 情報理工学研究科
情報・ネットワーク工学専攻
〒 182−8585 東京都調布市調布ヶ丘 1–5–1

3 提案手法: 相対関係を考慮した木間類似度

前章で指摘した通り，pq-gram は元の木における位置情報が失われている．そのため，pq-gram の順序が異なる木を類似していると判定するリスクがある．

そこで本研究では，pq-gram の木内の位置を表現することで類似検索の精度向上を試みる．単純には深さや兄弟ノード間の順番など，pq-gram の絶対位置を付与する手法が考えられるが，厳密な指定によりロバスト性に欠けた表現となることが危惧される．そのため，2 つの pq-gram のペアを取り出し，その相対関係を付与する手法を提案する．

具体的には，以下に示す手順により木 T を「pq-gram 2 つとその相対関係」からなる要素の集合 P として表し，その類似度を計算する．

1. T から pq-gram 集合 S を取り出す．
2. S 内の要素 t の全ペア $((t_i,\ t_j)|i,\ j \in N,\ 1 \leq i \neq j \leq |S|)$ を取り出す．
3. 取り出した各ペア $(t_i,\ t_j)$ に対し，相対関係 x を合わせた新たな要素 $(t_i,\ t_j,\ x_{t_i,\ t_j})$ からなる集合 $P = \{(t_i,\ t_j,\ x_{t_i,\ t_j})|1 \leq i \neq j \leq |S|\}$ を取り出す．

相対関係の決定方法については後述する．また，これ以降「pq-gram 2 つとその相対関係」の集合 P を単に pq-gram ペア集合 P と呼ぶ．pq-gram ペア集合 P_1，P_2 間の pq-gram ペア類似度 PSim は式 (3) のように Jaccard 係数で定義する．

$$\mathrm{PSim}(P_1,\ P_2) = \frac{|P_1 \bigcap P_2|}{|P_1 \bigcup P_2|} \tag{3}$$

3.1 相対関係の決定

x には T において t_1 が t_2 からみてどこにあるか，具体的には

- 2 つの pq-gram が祖先・子孫の関係にある場合，どちらが上にあるか (上下関係)
- 祖先・子孫の関係でない場合は，2 つの pq-gram のどちらが左にあるか (左右関係)

という関係を 1 つ入れる．これらの関係には pq-gram の q 個の子ノードを持つノードに着目し，本研究ではこれを中心ノードと呼ぶ．t_1，t_2 の中心ノードに対して LCA(最小共通祖先) を求め，その位置によって以下の 3 通りで場合分けし，相対関係を判断する．なお，LCA とは木中のある 2 ノードに共通する祖先のうち，根ノードから最も遠い位置にあるノードのことで，対象のノード自身が LCA となる場合も存在する．

関係 1 t_1，t_2 の中心ノードが一致する場合，左右関係を比較

関係 2 どちらか一方の中心ノードと LCA が一致する場合，上下関係を比較

関係 3 どちらの中心ノードも LCA と一致しない場合，左右関係を比較

3 つの関係について，$p = 1$, $q = 2$ として説明する．

図 4 は関係 1 の状況を表しており，t_1，t_2 の中心ノードは a で一致する．この場合，2 つの中心ノードの LCA も a で中心ノードと一致するため，中心ノードの一番左の子ノードが，t_1，t_2 のどちらに含まれているかによって左右を判断する．図 4 ではノード a の一番左の子ノード b は t_1 に含まれているため，t_1 は t_2 より左に存在すると判断でき，(t_1，t_2，左) となる．なお，兄弟関係にあるノードの左右は，木 T を前置順に探索した際に先に到達したノードが左に存在すると判断する．

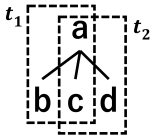

図 4 相対関係の関係例 1

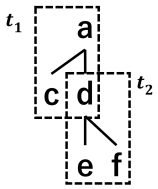

図 5 相対関係の関係例 2

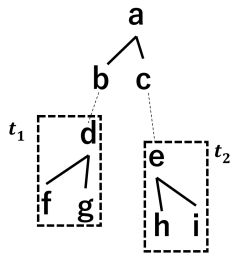

図 6 相対関係の関係例 3

図 5 は関係 2 の状況を表しており，t_1 の中心ノードは a，t_2 の中心ノードは d である．その LCA は a で，t_1 の中心ノードと一致する．どちらかの中心ノードが LCA である場合，LCA が含まれているほうの pq-gram がより上に存在すると判断できるため，(t_1，t_2，上) となる．

図 6 は関係 3 の状況を表しており，t_1 の中心ノードは d，t_2 の中心ノードは e である．一方，その LCA は t_1，t_2 どちらにも含まれていないノード a である．この場合，t_1，t_2 の各中心ノードの祖先のうち，LCA の子ノードの兄弟関係を比較する．図 6 では t_1 からノード b，t_2 からノード c が選ばれる．b は c より左にあると判断されるため，t_1 は t_2 より左に存在すると判断でき，(t_1，t_2，左) となる．

以上の判断基準を用い，2 つの pq-gram t_1，t_2 における相対関係決定アルゴリズムを Algolithm1 に示す．

4 実験

RNA STRAND[*1]の RNA データを 11 個利用し，各データについて以下の実験を行った．

4.1 データ作成

1 つの RNA を表す順序あり木を T とする．T に対して，以下の手順で部分木移動操作を適用した 10 個の木 T_1，T_2，$\cdots$，T_{10} を作る．

1. 根ノード以外のノードをランダムに一つ選択し，そのノードを根ノードとする部分木を st とする．
2. st を移動させる．移動先は st に含まれないノード v をランダムに一つ選択し，その子ノードとして st を挿入する．何番目の子として挿入するかはランダムに選択する．

[*1] http://www.rnasoft.ca/strand/

Algorithm 1 相対関係の決定 *compare*

Input: pq-gram t_1, t_2
Output: t_1 の t_2 から見た上下左右の位置

```
c1 ← t1 の中心ノード
c2 ← t2 の中心ノード
if c1 = c2 then
    if c1 の一番左の子ノードが t1 に含まれる then
        return left
    else
        return right
    end if
else
    lca ← c1と c2 の LCA
    if lca = c1 then
        return top
    else if lca = c2 then
        return under
    else
        clca1 ← c1 の祖先かつ lca の子ノード
        clca2 ← c2 の祖先かつ lca の子ノード
        if clca1 が clca2 より左に位置する then
            return left
        else
            return right
        end if
    end if
end if
```

上記のように部分木を 1 回移動する度に木を生成する．連続して計 10 回の部分木移動操作を行うことで，T_1, T_2, $\cdots$, T_{10} が生成される．さらにこれを 10 セット繰り返し，木 T を修正した合計 100 個の木の集合 F を作成した．

4.2 木編集距離との相関

T 及び F に含まれる全ての木 $\forall T_F \in F$ から pq-gram 集合 $S(T)$, $S(T_F)$ と pq-gram ペア集合 $P(T)$, $P(T_F)$ を取り出した．なお，pq-gram の大きさは $p = 1$, $q = 2$ とした．

そして $\forall T_F \in F$ に対して

- pq-gram 類似度 $\mathrm{Sim}(S(T), S(T_F))$
- pq-gram ペア類似度 $\mathrm{PSim}(P(T), P(T_F))$

を計算した．加えて木編集距離 $\delta(T, T_F)$ を計算し，pq-gram 類似度 Sim と pq-gram ペア類似度 PSim の，木編集距離との相関係数を求めた．木編集距離を計算する際にはノード挿入，削除，置換の各コストは全て 1 とした．

相関係数は表 1 のようになった．pq-gram ペア類似度との相関は全データにおいて-0.9 を超え，pq-gram 類似度よりも pq-gram ペア類似度の方が強い相関となった．また，図 7 は ASE_00001 における T 及び T_F の pq-gram 類似度，pq-gram ペア類似度と木編集距離を図示したものである．ASE_00001 は 187 個のノードを持っているが，木編集距離が 100 以上の時，pq-gram 類似度は 0.8 以上を保っているのに対し，pq-gram ペア類似度は 0.6 以下まで低下している．この傾向からも，pq-gram ペアは pq-gram 単体よりも木の変化に敏感に対応できている．なお，表 2 は各 RNA における T のノード数と，T から生成された pq-gram ペア集合 $P(T)$ において上下関係，左右関係が割り振られた要素の割合である．

RNA 名	pq-gram (従来手法)	pq-gram ペア (提案手法)
ASE_00001	-0.733	-0.929
ASE_00003	-0.608	-0.911
ASE_00009	-0.680	-0.949
ASE_00011	-0.590	-0.910
ASE_00024	-0.460	-0.956
ASE_00034	-0.701	-0.953
ASE_00056	-0.654	-0.921
ASE_00068	-0.645	-0.929
ASE_00089	-0.850	-0.948
ASE_00103	-0.556	-0.937
ASE_00116	-0.595	-0.928

表 1　各類似度と木編集距離の相関係数

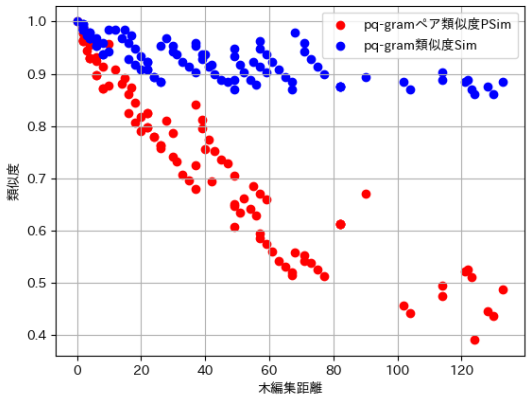

図 7　ASE_00001 における各類似度と木編集距離の散布図

RNA 名	上下関係 (%)	左右関係 (%)	ノード数
ASE_00001	57.3	42.7	187
ASE_00003	52.2	47.8	209
ASE_00009	65.2	34.8	162
ASE_00011	52.2	47.8	209
ASE_00024	53.1	46.9	173
ASE_00034	52.4	47.6	208
ASE_00056	53.1	46.9	177
ASE_00068	55.2	44.8	190
ASE_00089	57.4	42.6	172
ASE_00103	49.0	51.0	228
ASE_00116	55.4	44.6	186

表 2　各データのノード数及び pq-gram ペア間の相対関係の分布

4.3 Min-Hash を用いた近似類似度との相関

pq-gram ペア類似度 $\mathrm{PSim}(P(T), P(T_F))$ は pq-gram 類似度 $\mathrm{Sim}(S(T), S(T_F))$ よりも木編集距離との相関が強い．しかしペアを考えるため $P(T)$ の要素数が大きく，計算量の大きさが欠点である．この問題点に対処するため，Min-Hash による近似類似度を使うことを検

討した．Min-Hash は集合に対する Locally-Sensitive-Hashing であり，集合の要素を確率的な規則 π に従って，整数に写像し，写像後の最小値をハッシュ値とする．

Min-Hash のハッシュ関数 mh を使い，集合からハッシュ値に対応する要素を取り出した．まず，pq-gram 集合 S からは以下の手順に従い，pq-gram $e(S)$ を取り出した．

1. pq-gram 集合 S に対して π を適用し，その最小値を求める．
2. 最小値を実現する pq-gram $e(S)$ を取り出す．

また，pq-gram ペア集合 P からは以下の手順に従い pq-gram ペア $ep(S)$ を取り出した．

1. pq-gram 集合 S に対して π を適用し，最小値とその次に小さい値を求める．
2. 最小値とその次に小さい値を実現する pq-gram $e_1(S)$, $e_2(S)$ を取り出す．同じ形状の pq-gram が S に複数あった場合は，順序木から前置順に pq-gram を取り出した際に一番最後に取り出した pq-gram を取り出す．
3. $e_1(S)$, $e_2(S)$ の相対関係を求め，
   $ep(S) = (e_1(S), e_2(S),$ 相対関係) を取り出す．

加えて，100 個のハッシュ関数 $\mathrm{mh}_l(1 \leq l \leq 100)$ を用い，取り出した 100 個の要素の内いくつが一致していたかを計算することで，それぞれの集合に対する近似類似度を計算した．pq-gram 集合の近似類似度 MhSim は式 (4) で定義した．

$$\mathrm{MhSim}(S_1,\ S_2) = \frac{e_l(S_1) = e_l(S_2) \text{ となる } l \text{ の数}}{100} \quad (4)$$

pq-gram ペア集合 P_1, P_2 の pq-gram ペア近似類似度 MhPSim は式 (5) で定義した．

$$\mathrm{MhPSim}(P_1,\ P_2) = \frac{ep_l(S_1) = ep_l(S_2) \text{ となる } l \text{ の数}}{100} \quad (5)$$

以上の計算で得られた pq-gram 集合の近似類似度 MhSim と pq-gram ペア集合の近似類似度 MhPSim の，木編集距離との相関係数について評価した．

実験結果は表 3 のようになり，ハッシュ関数を利用しても，全データにおいて pq-gram 類似度よりも pq-gram ペア類似度の方が強い相関をとることが確認できた．一方で表 1 と比較すると，ハッシュ関数を用いたことで全体的に相関は弱まった．これは用いたハッシュ関数が原因の一つだと考えられる．現在用いているハッシュ関数では，多重集合に対して一定の pq-gram しか取り出せず，偏りが生じている．そのため，今後はハッシュ関数が多重集合を考慮できるように改善していく必要がある．

5 結論

木編集距離は計算量が大きいため，木を部分構造の集合として表現し，その集合間類似度を計算することで，木間類似度を近似する手法が研究されてきた．Peisen ら [4] や Nikolaus ら [1] は根あり順序あり木を pq-gram の集合として表現し，pq-gram 類似度を木間類似度として高速に計算した．しかし，この表現方法では pq-gram の順序が異なる木を類似していると判定することが起こりえる．

RNA データ名	pq-gram (従来手法)	pq-gram ペア (提案手法)
ASE_00001	-0.718	-0.890
ASE_00003	-0.274	-0.784
ASE_00009	-0.627	-0.914
ASE_00011	-0.275	-0.868
ASE_00024	-0.468	-0.934
ASE_00034	-0.615	-0.913
ASE_00056	-0.706	-0.914
ASE_00068	-0.601	-0.861
ASE_00089	-0.676	-0.899
ASE_00103	-0.688	-0.789
ASE_00116	-0.322	-0.702

表 3　ハッシュ関数を用いた場合の各近似類似度と木編集距離の相関係数

そこで本研究では，2 つの pq-gram とその相対関係を新たな集合とすることで，木内の位置を表現する手法を提案した．実験では既存手法の pq-gram 類似度と新たに提案した pq-gram ペアによる類似度の，木編集距離との相関係数を評価した．用いた全てのデータにおいて pq-gram ペア類似度の相関の方が強いという結果になり，提案手法の有意性が確認できた．

今後は提案手法の改善に取り組む予定である．pq-gram ペアのある関係に対して，ペアは実際にどれだけのノード数離れているのか，例えば上下関係にある pq-gram はどれだけの高低差があるのかといった，距離についても考慮することで，さらに相関を強めることができると考えられる．

謝辞

本研究は JSPS 科研費 JP21K11901 の助成を受けたものである．

参考文献

[1] N. Augsten, M. Böhlen, and J. Gamper. Approximate matching of hierarchical data using pq-grams. In *Proc. of the 31st VLDB*, pages 301–312, 2005.

[2] P. Bille. A survey on tree edit distance and related problems. *Theoretical Computer Science*, 337(1):217–239, 2005.

[3] A. Broder. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*, pages 21–29, 1997.

[4] P. Yuan, C. Sha, X. Wang, B. Yang, A. Zhou, and S. Yang. Xml structural similarity search using mapreduce. In *Proc. WAIM 2010*, pages 169–181. Springer, 2010.