



Title	研究紹介
Author(s)	平田, 耕一
Citation	2010年度科学技術振興機構ERATO湊離散構造処理系プロジェクト講究録. p.348-350.
Issue Date	2011-06
Doc URL	http://hdl.handle.net/2115/48377
Type	conference presentation
Note	ERATO 湊離散構造処理系プロジェクト春のワークショップ (キックオフシンポジウム). 2010年5月28日 (金) ~ 29日 (土). ERATO湊プロジェクト研究室.
File Information	26.hirata_06.pdf



[Instructions for use](#)

研究紹介

平田耕一
九州工業大学情報工学研究院

大学の位置



大学のあゆみ

- 明治40(1907)年 明治専門学校(4年制)設立(2年後開校)
- 大正10(1921)年 官立明治専門学校へ移管(4年制)
- 昭和19(1944)年 明治工業専門学校と改称(3年制)
- 昭和24(1949)年 九州工業大学設置
- 昭和40(1965)年 工学研究科修士課程設置
- 昭和61(1986)年 情報工学部設置(学生受入は翌年から)
- 昭和63(1988)年 工学研究科博士課程設置
- 平成 3(1991)年 情報工学研究科修士課程設置
- 平成 5(1993)年 情報工学研究科博士課程設置
- 平成12(2000)年 生命体工学研究科博士課程設置
- 平成16(2004)年 国立大学法人 九州工業大学

離散構造

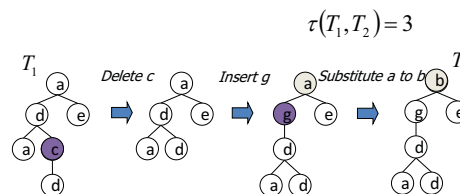
- 文字列
- 木
- グラフ
- 超グラフ
-

Distance measures for trees

- 共通部分最大化
 - 似ている部分が大きければ似ている
 - 木の編集距離
- 共通部分頻度最大化
 - 似ている部分が高頻度であれば似ている
 - 局所頻度距離
- 文字列編集距離
 - 木を文字列で表現した後の文字列編集距離
 - オイラー文字列, 二分木符号

Tree edit distance

- 木の編集距離 $\tau(T_1, T_2)$ [Tai 79]
- 木から木へ変換する編集操作の最小数
 - 代入
 - 削除
 - 挿入



Tree edit distance

時間計算量

- n : the maximum number of nodes
- $O(n^6)$ [Tai 79]
- $O(n^4)$ [Zhang & Shasha 89]
- $O(n^3 \log n)$ [Klien 98]
- $O(n^3)$ [Demaine et al. 06]

8

Local frequency distance

共通部分頻度最大化

- Leaf, degree and label histograms [Kailing et al. 04]
- Binary branch [Yang et al. 05]
- q-gram [Kuboyama et al. 06], bifoliate q-gram [Kuboyama et al. 08]
- pq-gram [Augsten et al. 05]
- Sibling histogram [Aratsu et al, 08]
- Efficient but not metric in general
- (Some of them give) constant factor lower bound on tree edit distance

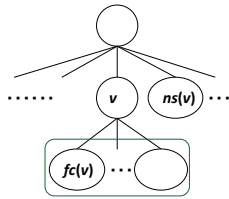
9

Local frequency distance

局所情報の組合せ

Local label

- v : a node
- $ns(v)$: the next sibling of v
- $fc(v)$: the first child of v
- $ch(v)$: the children of v

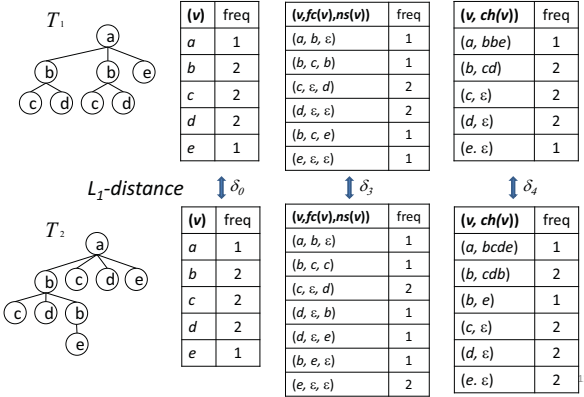


f_i	Local information
f_0	(v)
f_1	$(v, fc(v))$
f_2	$(v, ns(v))$
f_3	$(v, fc(v), ns(v))$
f_4	$(v, ch(v))$

f_0 : label histogram [Kailing 04]
 f_3 : binary branch [Yang 05]
 f_4 : sibling histogram [Aratsu 08]

10

Local frequency distance



12

Local frequency distance

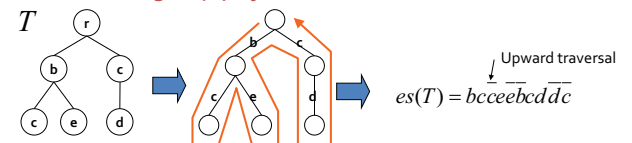
Distances	Time complexity	Lower bound (τ : ted)	
Tree edit distance	$O(n^3)$	-	
Leaf histogram [Kailing et al. 04]	$O(n)$	$\leq \tau$	
Degree histogram [Kailing et al. 04]	$O(n)$	$\leq 3\tau$	
Label histogram [Kailing et al. 04]	$O(n)$	$\leq 2\tau$	δ_0
Binary branch [Yang et al. 05]	$O(n)$	$\leq 5\tau$	δ_3
q-gram [Kuboyama et al. 06]	$O(qn)$	(not exist)	
Bifoliate q-gram [Kuboyama et al. 08]	$O(g \min(q, d) \ln)$	(not exist)	
pq-gram [Augsten et al. 05]	$O(pqn)$	(not exist)	
Sibling histogram [Aratsu et al. 08]	$O(n)$	$\leq 4\tau$	δ_4

l : the number of leaves, d : depth, g : degree

13

String representation of trees

Euler string $es(T)$ of a tree T



σ : string edit distance

[Akutsu 06]

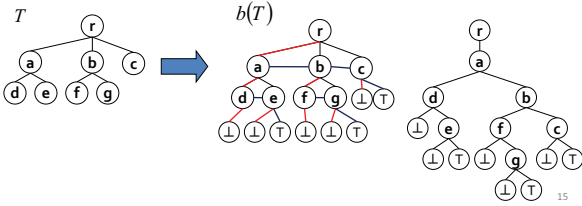
$$\frac{\sigma(es(T_1), es(T_2))}{2} \leq \tau(T_1, T_2) \leq (2h+1)\sigma(es(T_1), es(T_2))$$

h : minimum height

14

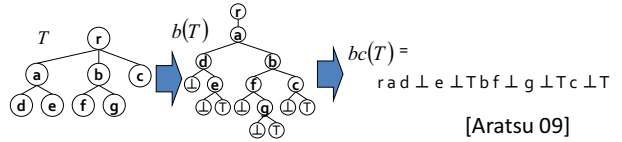
String representation of trees

- **Binary tree representation $b(T)$ of a tree T**
 - $fc(v)$ (or \perp if no exists) is a left child of v in $b(T)$
 - $ns(v)$ (or \top if no exists) is a right child of v in $b(T)$
 - If v is a root r , then r has just a left child



String representation of trees

- **Binary tree code $bc(T)$ of a tree T**
 - The string obtained by preorder traversal of $b(T)$



[Aratsu 09]

$$\frac{\sigma(bc(T_1), bc(T_2))}{2} \leq \tau(T_1, T_2) \leq (h+1)\sigma(bc(T_1), bc(T_2)) + h$$

16

Distance measures for trees

	Measure	Complexity	Metric	編集距離の近似
木の編集距離	共通部分のサイズ	$O(n^3)$	Yes	
木の編集距離の亜種 top-down, bottom-up, degree-2, constrained	共通部分のサイズ	$O(n^2)$	Yes	定数上界
局所頻度距離	共通部分の頻度	$O(n)$	No	定数下界
文字列表現	文字列編集距離	$O(n^2)$	Yes	定数下界・上界
Edit Sensitive Parsing [Garofalakis 05]	共通部分の頻度 +埋め込み	$O(n \log^* n)$	Yes	下界・定数上界 (移動付き)
今の興味		$O(n)$ $\sim O(n \log n)$	Yes	定数下界・上界

17