2021.02.03 冬のLAシンポジウム2020

ラベル付き木に対する索引構造

九州大学　稲永俊介

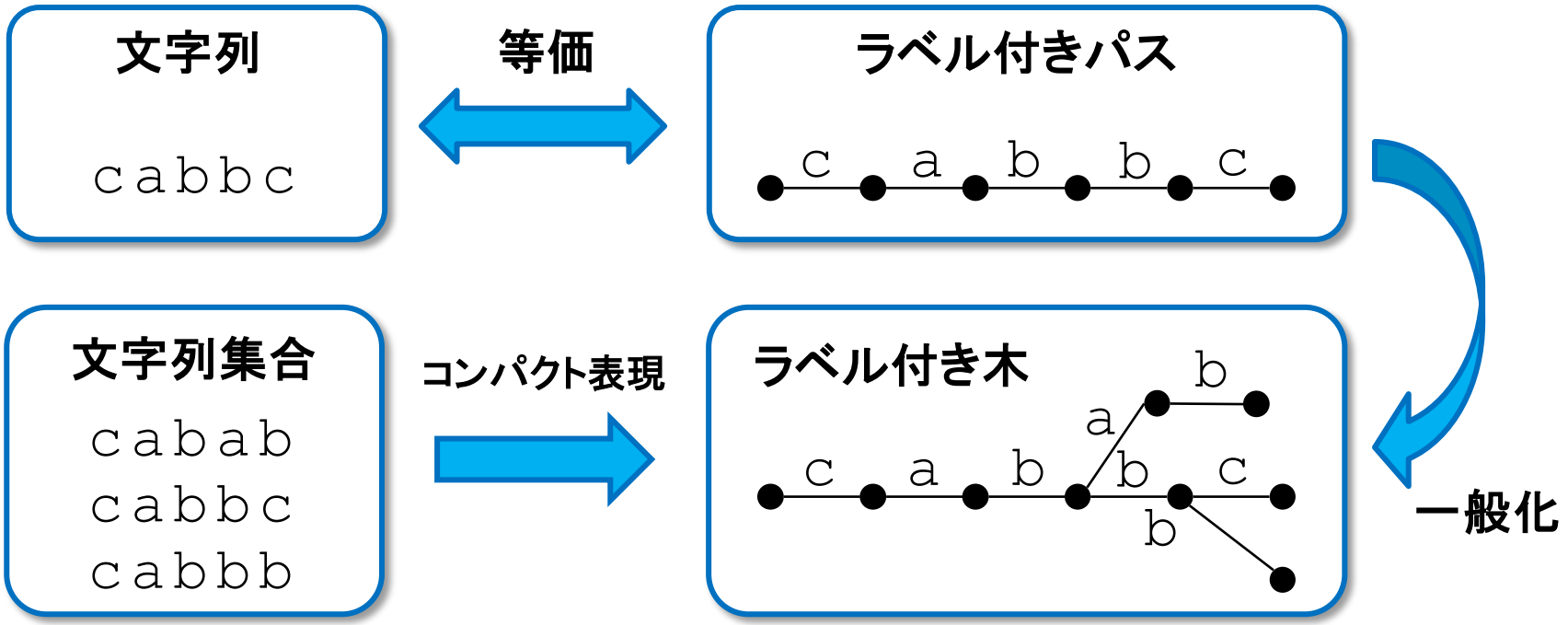
Towards a complete perspective on labeled tree indexing: new size bounds, efficient constructions, and beyond

情報処理学会 60周年記念論文

ラベル付き木

記号の連鎖である**文字列**は，
辺を文字でラベル付けしたパスと見なせる．

ラベル付き木は，文字列の自然な一般化であり，
また文字列集合のコンパクトな表現でもある．



ラベル付き木の索引問題

ラベル付き木上の部分パス照合クエリのための
データ構造(索引)のサイズとその構築方法を考える.

問題

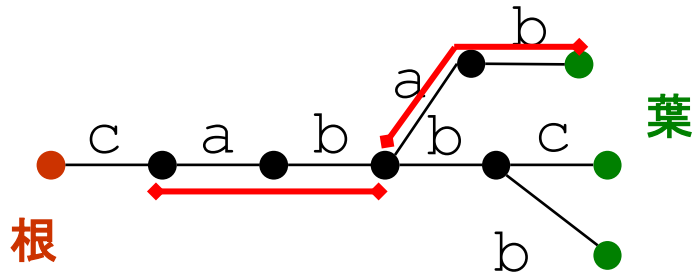
前処理入力: ラベル付き木 $\mathcal{T}$
クエリ入力: パターン文字列 P
クエリ出力: P と合致する $\mathcal{T}$ のすべての部分パス

パス文字列の読み方を2通り考える:

- **順木** $\mathcal{T}$: パスを **根** から **葉** に向かって読む.
- **逆木** $\mathcal{T}^R$: パスを **葉** から **根** に向かって読む.

順木/逆木の索引問題

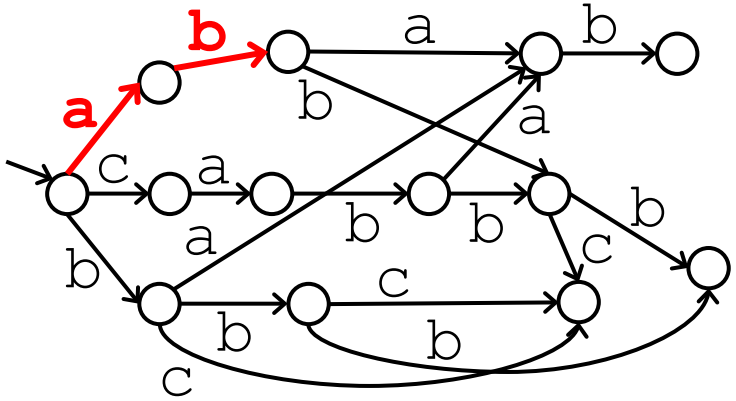
順木$\mathcal{T}$



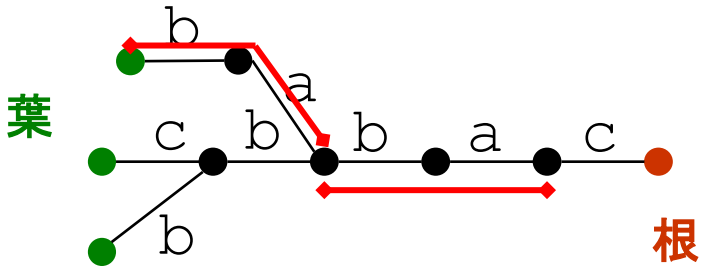
パターン $P = \mathtt{ab}$

DAWG($\mathcal{T}$)

$\mathcal{T}$の部分文字列を受理するオートマトン



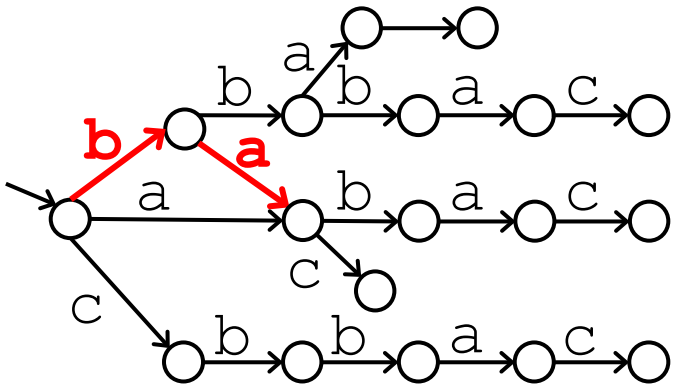
逆木$\mathcal{T}^R$



パターン $P^R = \mathtt{ba}$

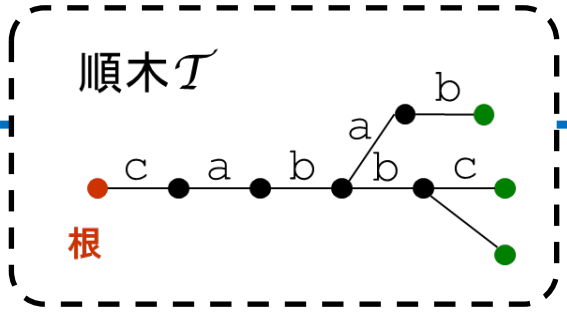
DAWG($\mathcal{T}^R$)

$\mathcal{T}^R$の部分文字列を受理するオートマトン



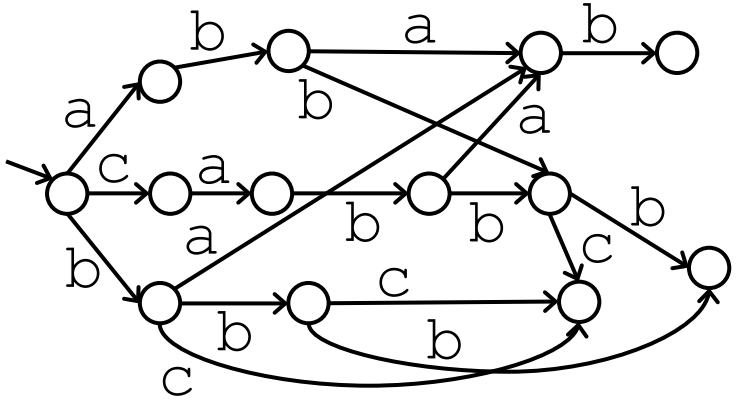
順木に対する索引構造

順木$\mathcal{T}$に対する
3種類の索引構造



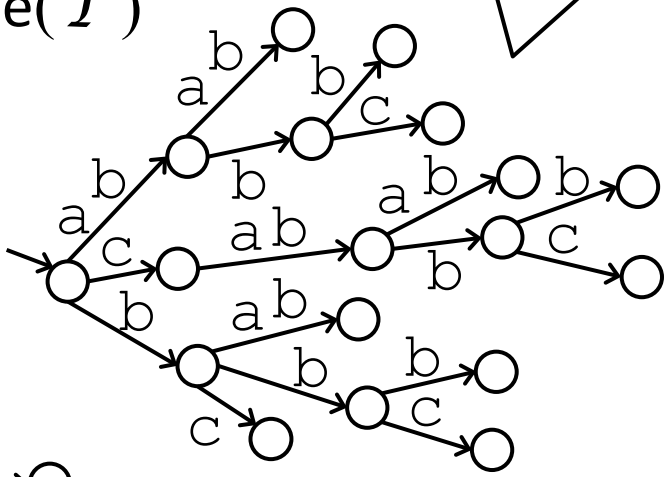
DAWG($\mathcal{T}$)

$\mathcal{T}$の部分文字列を
受理するオートマトン



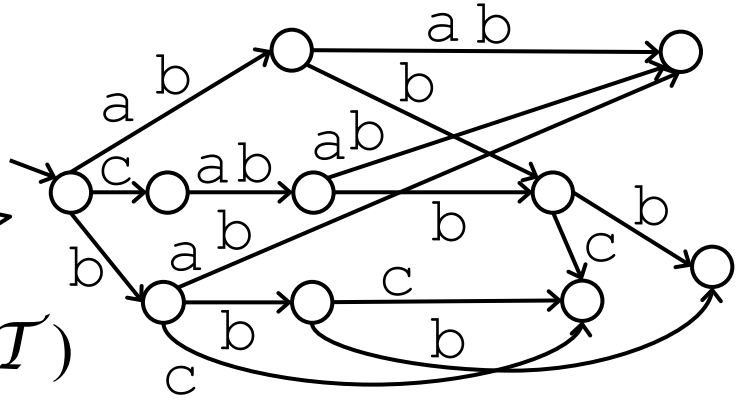
SuffixTree($\mathcal{T}$)

$\mathcal{T}$の部分文字列を
表現するコンパクト木



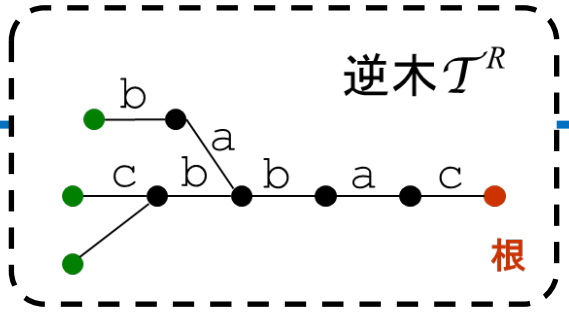
$\mathcal{T}$の部分文字列を
表現するコンパクトグラフ

CDAWG($\mathcal{T}$)



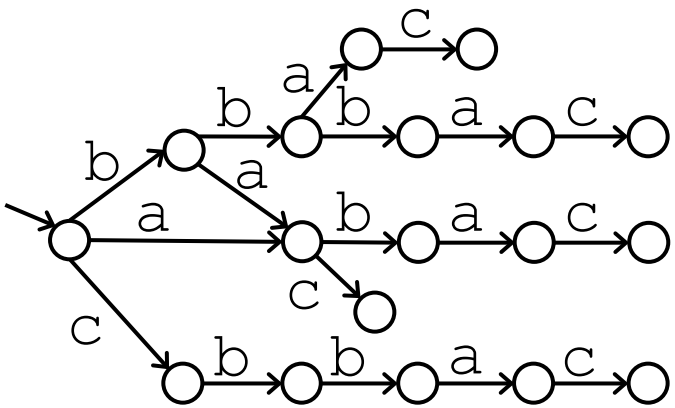
逆木に対する索引構造

逆木$\mathcal{T}^R$に対する
3種類の索引構造



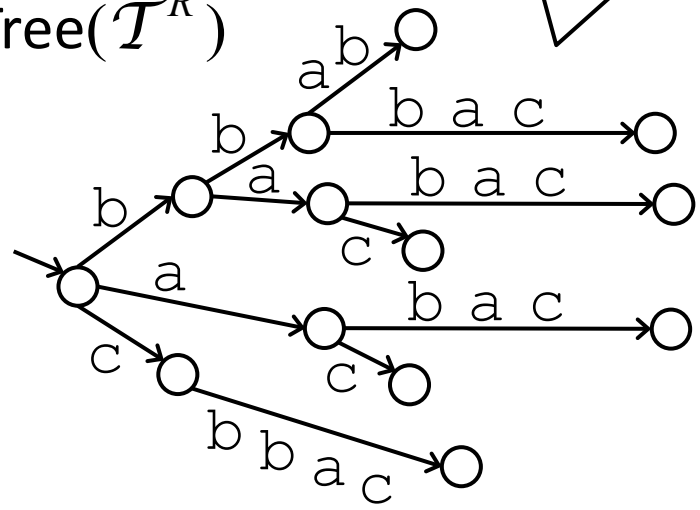
DAWG($\mathcal{T}^R$)

$\mathcal{T}^R$の部分文字列を
受理するオートマトン



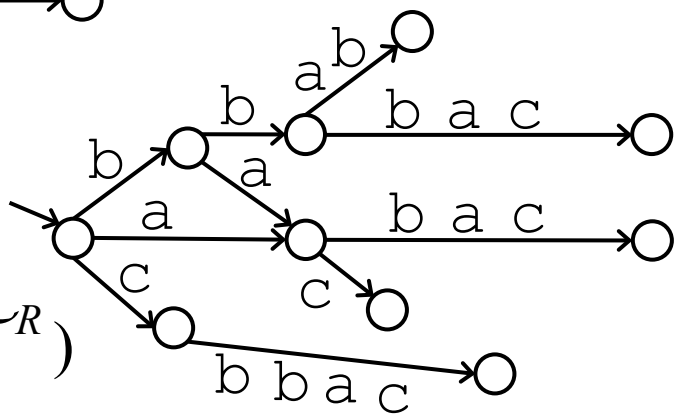
SuffixTree($\mathcal{T}^R$)

$\mathcal{T}^R$の部分文字列を
表現するコンパクト木



$\mathcal{T}^R$の部分文字列を
表現するコンパクトグラフ

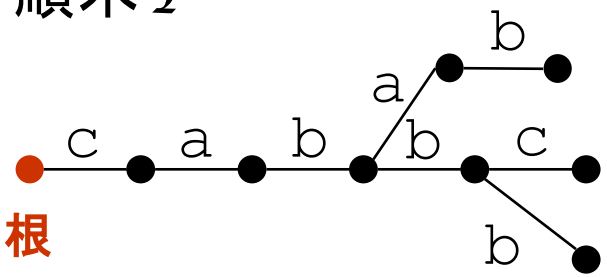
CDAWG($\mathcal{T}^R$)



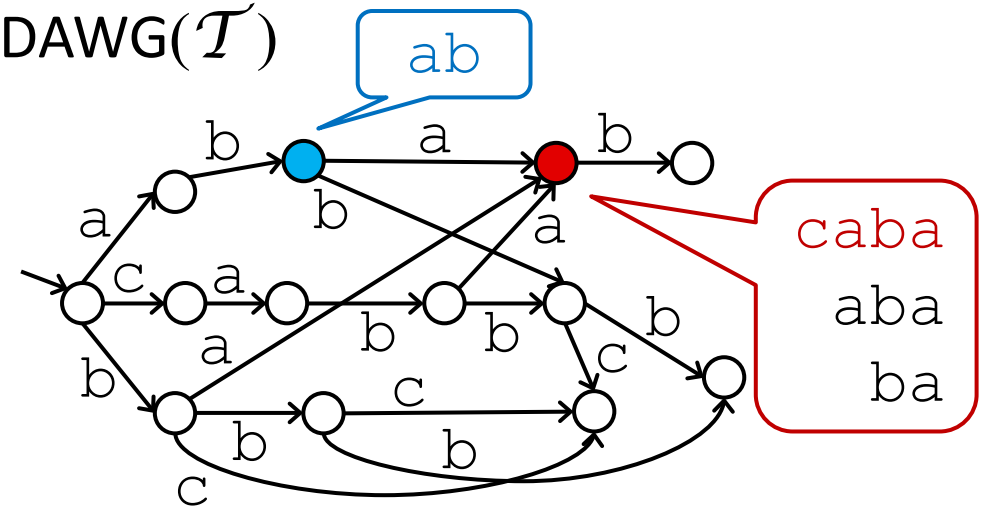
DAWG (Directed Acyclic Word Graph)

順木$\mathcal{T}$の部分文字列（部分パス）Xが **左極大** である ⇔
(1) aXとbXが$\mathcal{T}$の部分文字列である文字 a, b $(a \neq b)$ が存在, or
(2) Xは順木$\mathcal{T}$の**根**で始まる出現を持つ.

順木$\mathcal{T}$



DAWG($\mathcal{T}$)

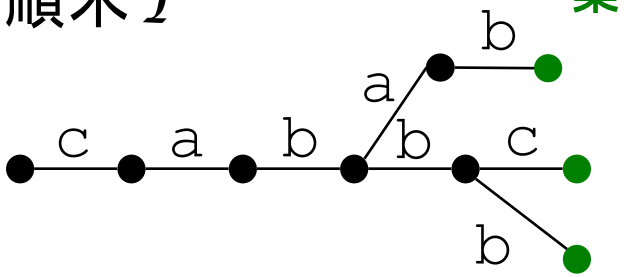


- 逆木$\mathcal{T}^R$ に対する DAWG も同様に定義.
- 文字列に対する DAWG [Blumer et al.] をラベル付き木に拡張.

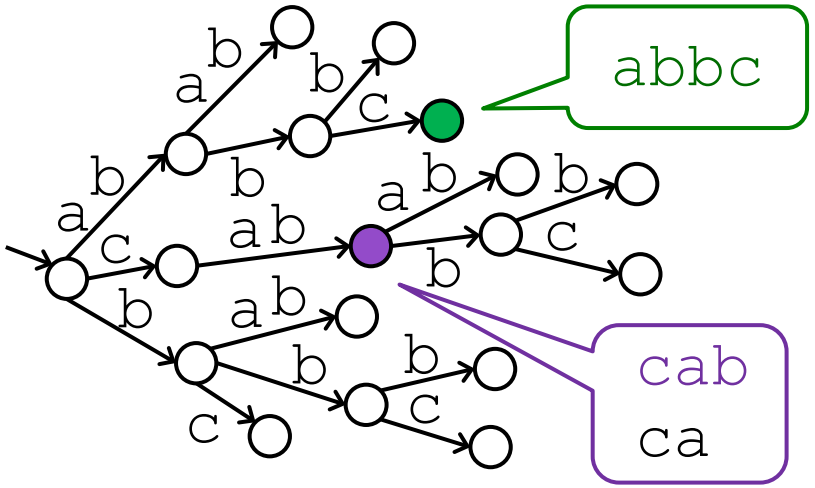
Suffix Trees（接尾辞木）

順木$\mathcal{T}$の部分文字列（部分パス）Xが **右極大** である ⇔
(1) XaとXbが$\mathcal{T}$の部分文字列である文字 a, b $(a \neq b)$ が存在, or
(2) Xは順木$\mathcal{T}$の**葉**で終わる出現を持つ.

順木$\mathcal{T}$



SuffixTree($\mathcal{T}$)

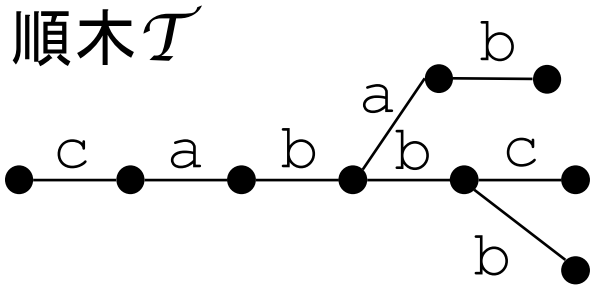


- 逆木$\mathcal{T}^R$ に対する Suffix Tree も同様に定義.
- 文字列に対する Suffix Tree [Weiner] をラベル付き木に拡張.

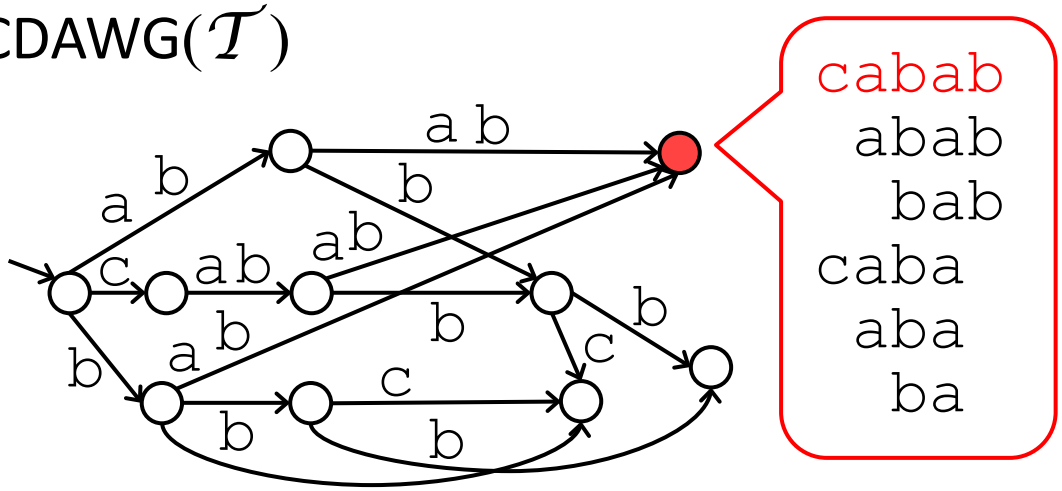
CDAWG (Compact DAWG)

順木$\mathcal{T}$の部分文字列(部分パス)Xが **両極大** である ⇔
Xは順木$\mathcal{T}$上で左極大かつ右極大である.

順木$\mathcal{T}$



CDAWG($\mathcal{T}$)



- 直感:CDAWG は DAWG と Suffix Tree のハイブリッドデータ構造.
- 逆木$\mathcal{T}^R$ に対する CDAWG も同様に定義.
- 文字列に対する CDAWG [Blumer ら] をラベル付き木に拡張.

ラベル付き木に対する索引構造のサイズ

既存研究 [Mohri et al. / Kosaraju / Kimura & Kashima]

上界

	順木$\mathcal{T}$		逆木$\mathcal{T}^R$	
索引	頂点数	辺数	頂点数	辺数
DAWG	$2n-3$	–	–	–
CDAWG	–	–	–	–
Suffix Tree	–	–	$2n-3$	$2n-4$
Suffix Array	–		$n+1$	

n: 入力順木/逆木の頂点数

ラベル付き木に対する索引構造のサイズ

本研究

上界

	順木$\mathcal{T}$		逆木$\mathcal{T}^R$	
索引	頂点数	辺数	頂点数	辺数
DAWG	$2n-3$	$\mathbf{O(n^2)}$	$\mathbf{O(n^2)}$	$\mathbf{O(n^2)}$
CDAWG	$\mathbf{2n-3}$	$\mathbf{O(n^2)}$	$\mathbf{2n-3}$	$\mathbf{2n-4}$
Suffix Tree	$\mathbf{O(n^2)}$	$\mathbf{O(n^2)}$	$2n-3$	$2n-4$
Suffix Array	$\mathbf{O(n^2)}$		$n+1$	

n: 入力順木/逆木の頂点数

※ 文字列（パス木）の場合はすべて $O(n)$

索引サイズのタイトな上界・下界を証明

本研究

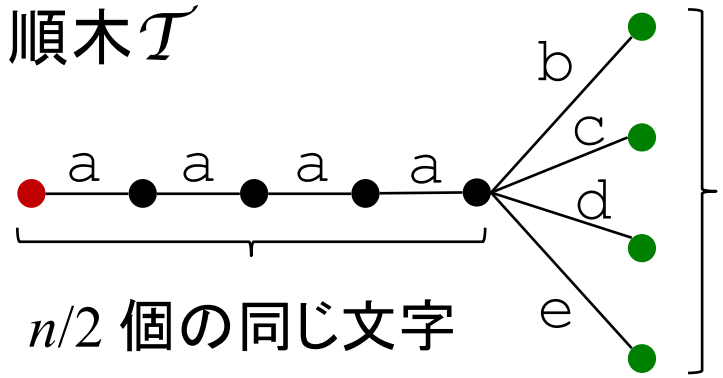
上界

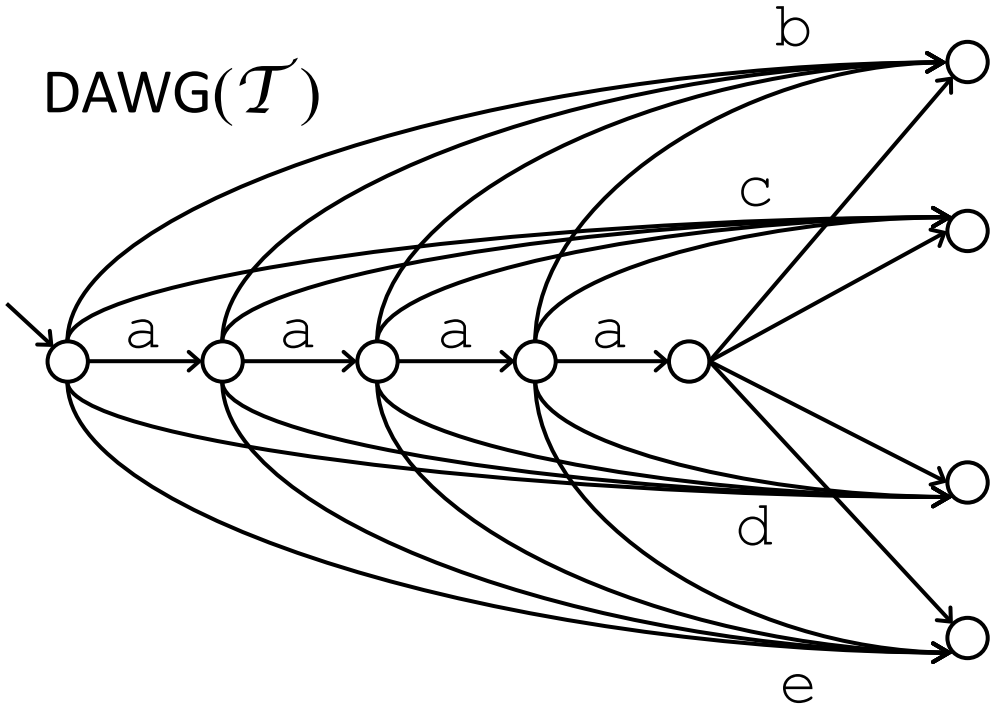
	順木$\mathcal{T}$		逆木$\mathcal{T}^R$	
索引	頂点数	辺数	頂点数	辺数
DAWG	$2n-3$	$\mathbf{O(n^2)}$	$\mathbf{O(n^2)}$	$\mathbf{O(n^2)}$
CDAWG	$\mathbf{2n-3}$	$\mathbf{O(n^2)}$	$\mathbf{2n-3}$	$\mathbf{2n-4}$
Suffix Tree	$\mathbf{O(n^2)}$	$\mathbf{O(n^2)}$	$2n-3$	$2n-4$
Suffix Array	$\mathbf{O(n^2)}$		$n+1$	

下界

	順木$\mathcal{T}$		逆木$\mathcal{T}^R$	
索引	頂点数	辺数	頂点数	辺数
DAWG	$\mathbf{2n-3}$	$\mathbf{\Omega(n^2)}$	$\mathbf{\Omega(n^2)}$	$\mathbf{\Omega(n^2)}$
CDAWG	$\mathbf{2n-3}$	$\mathbf{\Omega(n^2)}$	$\mathbf{2n-3}$	$\mathbf{2n-4}$
Suffix Tree	$\mathbf{\Omega(n^2)}$	$\mathbf{\Omega(n^2)}$	$\mathbf{2n-3}$	$\mathbf{2n-4}$
Suffix Array	$\mathbf{\Omega(n^2)}$		$n+1$	

順木に対する DAWG の辺数の下界





辺数 : $n/2 \times n/2 = \Omega(n^2)$

順木$\mathcal{T}$に対する線形領域索引は存在するか？

	順木$\mathcal{T}$		逆木$\mathcal{T}^R$	
索引	頂点数	辺数	頂点数	辺数
DAWG	$2n-3$	$\Theta(n^2)$	$\Theta(n^2)$	$\Theta(n^2)$
CDAWG	$2n-3$	$\Theta(n^2)$	$2n-3$	$2n-4$
Suffix Tree	$\Theta(n^2)$	$\Theta(n^2)$	$2n-3$	$2n-4$
Suffix Array	$\Theta(n^2)$		$n+1$	

n: 入力順木/逆木の頂点数

【難しさ】辺をポインタで明示的に保持すると，
最悪時にはどう頑張っても $\Theta(n^2)$ 領域必要．

簡潔データ構造を用いたとしても $\Omega(n^2)$ ビット必要．

順木$\mathcal{T}$に対する線形領域索引は存在するか？

	順木$\mathcal{T}$		逆木$\mathcal{T}^R$	
索引	頂点数	辺数	頂点数	辺数
DAWG	$2n-3$	$\Theta(n^2)$	$\Theta(n^2)$	$\Theta(n^2)$
CDAWG	$2n-3$	$\Theta(n^2)$	$2n-3$	$2n-4$
Suffix Tree	$\Theta(n^2)$	$\Theta(n^2)$	$2n-3$	$2n-4$
Suffix Array	$\Theta(n^2)$		$n+1$	

Yes!

n: 入力順木/逆木の頂点数

定理 [本研究]

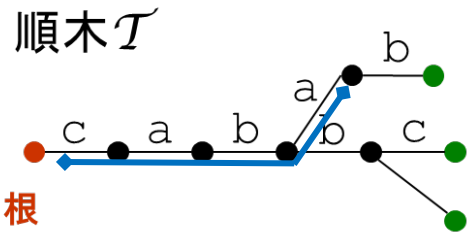
順木$\mathcal{T}$に対するDAWGの**<u>$O(n)$ 領域コンパクト表現</u>**が存在し，それを$O(n)$時間で構築できる．

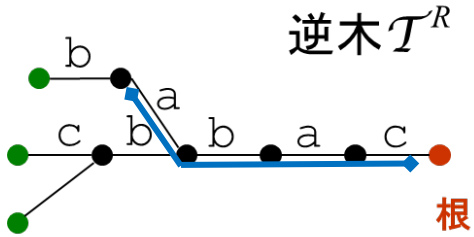
このコンパクト表現を用いて**<u>木上の双方向パターン照合</u>**を$O(m \log \sigma + occ)$時間で実行できる．

n: 入力木の頂点数，m: パターン長，σ: アルファベットサイズ，occ: パターンの出現回数

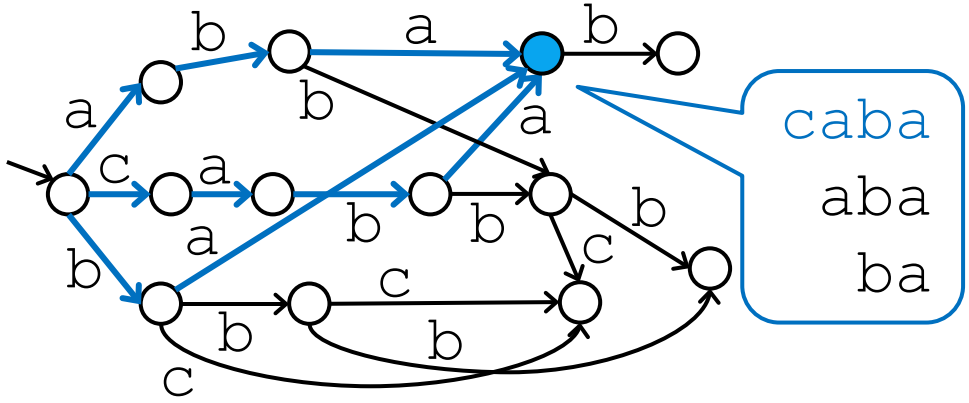
DAWG($\mathcal{T}$) と SuffixTree($\mathcal{T}^R$) の関係

順木$\mathcal{T}$で左極大　⇔　逆木$\mathcal{T}^R$で右極大

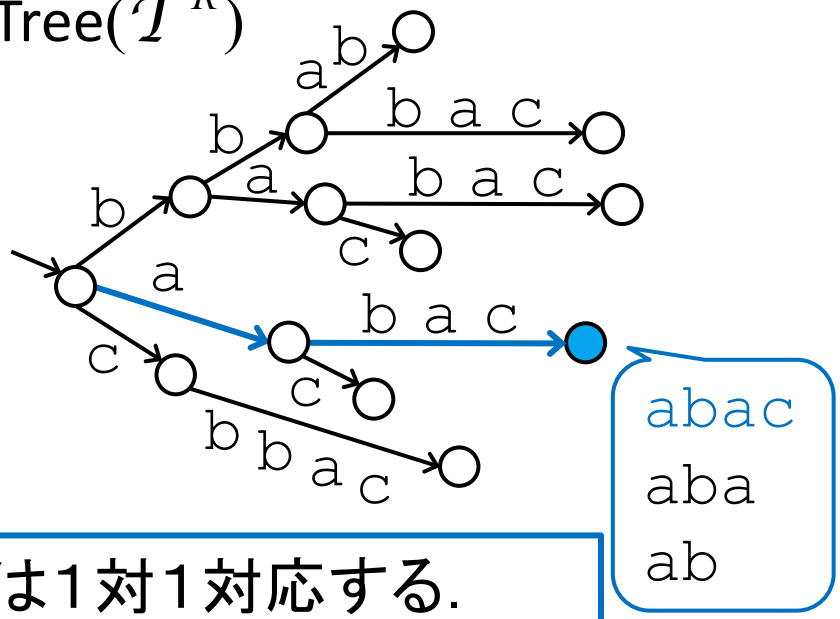




DAWG($\mathcal{T}$)



SuffixTree($\mathcal{T}^R$)

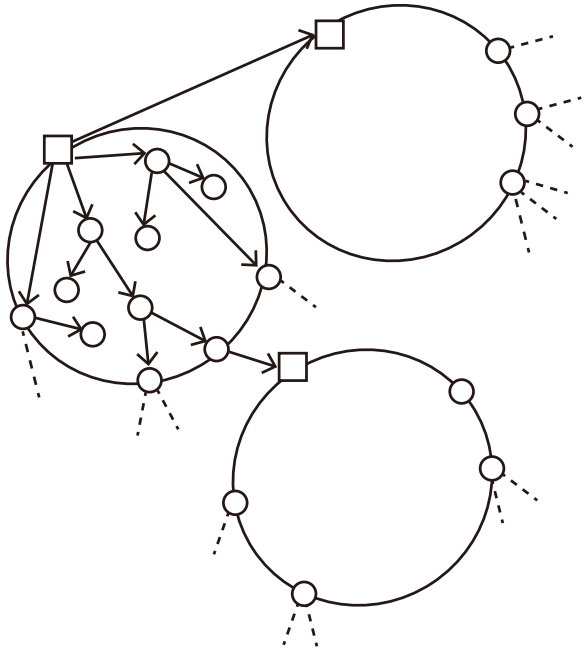


DAWG($\mathcal{T}$) と SuffixTree($\mathcal{T}^R$)の頂点は1対1対応する.
しかし辺には同様の対応がない（DAWG の辺は $O(n^2)$ 個）

順木$\mathcal{T}$に対する線形領域索引

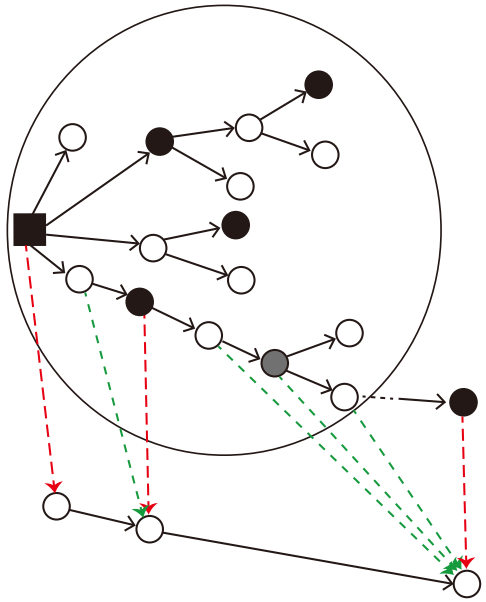
DAWG($\mathcal{T}$) の $O(n^2)$ 個あるすべての辺を，
SuffixTree($\mathcal{T}^R$) 上で $O(n)$ 領域だけを使ってシミュレート

SuffixTree($\mathcal{T}^R$) をサイズ $O(\sigma)$ の $O(n/\sigma)$ 個のクラスタに分解する



σ はアルファベットサイズ

各クラスタ中に，**適切に選んだ** DAWGの辺を陽に保持しておき，残りのDAWG 辺は組合せ的性質を利用して逐次的に復元する



まとめと未解決問題

- ラベル付き木に対する索引構造のサイズに関する<u>タイトな上界と下界</u>を与えた.

	順木$\mathcal{T}$		逆木$\mathcal{T}^R$	
索引	頂点数	辺数	頂点数	辺数
DAWG	$2n-3$	$\Theta(n^2)$	$\Theta(n^2)$	$\Theta(n^2)$
CDAWG	$2n-3$	$\Theta(n^2)$	$2n-3$	$2n-4$
Suffix Tree	$\Theta(n^2)$	$\Theta(n^2)$	$2n-3$	$2n-4$
Suffix Array	$\Theta(n^2)$		$n+1$	

- 順木に対する DAWG の $O(n)$ 領域表現を与えた.
  ➔ <u>**木上の双方向パターン照合を$O(n)$領域**</u>で初実現.

- 順木に対する CDAWG の $O(n)$ 領域表現は存在するか？